THE EVOLVING FUZZY CLUSTERING APPROACH FOR DISCRIMINATING NEUTRON AND GAMMA-RAY PULSES

ALI SEYED SHIRKHORSHIDI

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITI MALAYA
KUALA LUMPUR

2020

EVOLVING FUZZY CLUSTERING APPROACH FOR DISCRIMINATING NEUTRON AND GAMMA-RAY PULSES

ALI SEYED SHIRKHORSHIDI

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INFORMATION SYSTEMS, FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITI MALAYA
KUALA LUMPUR

2020

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **ALI SEYED SHIRKHORSHIDI**

Registration/Matric No: **WHA130003**

Name of Degree: **DOCTOR OF PHILOSOPHY**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

EVOLVING FUZZY CLUSTERING APPROACH FOR DISCRIMINATING NEUTRON AND GAMMA-RAY PULSES

Field of Study: **DATA MINING** (**TIME-SERIES CLUSTERING)**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyrighted work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyrighted work;
(5) I hereby assign all and every right in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ABSTRACT

Having a significant amount of data is not useful unless the data can be processed for extracting knowledge and information. One of the elementary steps in crunching data is to break it down into groups. When the data is small and collected in a controlled manner, and when the training data is appropriately labelled, the trivial approach is to use supervised learning to perform the grouping. Supervised methods need training data and information about groups beforehand; however, in the current reality, with an avalanche of data, this information is not available. Nevertheless, the need for grouping data remains. Clustering, as an unsupervised method, helps in these situations to group the data. However, unsupervised methods are usually less accurate than their supervised counterparts. To solve this drawback, unsupervised methods are often used as a pre-processing step, along with human judgment, to prune the data to create a reliable training set for the supervised process. One reason that clustering approaches do not yield desirable accuracy is that they will attempt to perform the procedure on all data, which may contain noise or outliers, and they do not have any mechanism by which to set aside the problematic data. Pulse-shape discrimination (PSD) for neutron and gamma-ray pulses that is addressed in this research is one example of a real-world case study that faces the same issues. Although the data utilised for this study is from a liquid scintillator, it can be applied to other signal detectors as well. Aside from this particular dataset, the proposed approach has been applied to a set of publicly available multivariate and time series datasets to prove the performance of the presented approach through an exploratory study. The evolving fuzzy clustering approach (EFCA) proposed in this study utilises a fuzzy membership matrix in fuzzy clustering to propose a new approach for clustering that embeds a heuristic post-pruning solution to address the aforementioned drawback. The method is an EFCA that attempts to find clusters of similar shapes with better

accuracy. It introduces an approach for post-pruning that is examined not only on neutron and gamma-ray discrimination but also on various datasets. The outcomes of the proposed method are evaluated against the traditional fuzzy *C*-means method and another well-known crisp clustering method, namely, *K*-means. For neutron and gamma-ray discrimination, the EFCA improved the Rand index (RI) accuracy by almost 8%. For other multivariate and time series datasets utilised in this study, results demonstrate the achievement of significant accuracy improvements for some of these datasets after heuristic post-pruning, resulting in 100% RI accuracy for some of them.

Keywords: Clustering, Fuzzy Clustering, Unsupervised Learning, Pre-processing, Pruning, Neutron and Gamma-ray discrimination

ABSTRAK

Mempunyai sejumlah besar data tidak berguna melainkan data boleh diproses untuk mengekstrak pengetahuan dan maklumat. Salah satu langkah asas dalam memproses data adalah untuk memecahnya ke dalam kumpulan. Apabila data kecil dan dikumpulkan secara terkawal, dan apabila data latihan dilabel dengan tepat, pendekatan remeh adalah menggunakan pembelajaran yang diawasi untuk melaksanakan kumpulan. Kaedah yang diselia memerlukan data latihan dan maklumat mengenai kumpulan terlebih dahulu; Walau bagaimanapun, dalam realiti semasa dengan data yang mencurah-curah, maklumat ini tidak tersedia. Namun, keperluan untuk mengelompokkan data adalah kekal. Gugusan sebagai kaedah yang tidak diselia membantu dalam situasi ini untuk mengelompokkan data. Walau bagaimanapun, kaedah yang tidak diselia biasanya kurang tepat berbanding kaunterpart yang diawasi mereka. Untuk menyelesaikan kelemahan ini, kaedah yang tidak diselia sering digunakan sebagai langkah pra-pemprosesan bersama-sama dengan pertimbangan manusia untuk mengurangkan data bagi mewujudkan satu set latihan yang boleh dipercayai untuk proses yang diselia. Salah satu sebab bahawa pendekatan gugusan tidak menghasilkan ketepatan yang diingini ialah mereka akan cuba melaksanakan prosedur pada semua data yang mungkin mengandungi bunyi atau nilai terpencil dan mereka tidak mempunyai mekanisme untuk mengetepikan data bermasalah. Diskriminasi bentuk denyut jantung (PSD) untuk denyut nadi neutron dan sinar gamma yang dikemukan dalam kajian ini adalah satu contoh kajian kes dunia sebenar yang menghadapi masalah yang sama. Walaupun data yang digunakan untuk kajian ini adalah daripada bahan kelip cecair, ia boleh digunakan untuk pengesan isyarat yang lain juga. Selain daripada set data tertentu, pendekatan yang dicadangkan telah digunakan untuk satu set multivariat awam yang tersedia dan set data siri masa untuk membuktikan prestasi pendekatan yang dibentangkan melalui kajian teroka. Pendekatan gugusan kabur yang ditambah baik (EFCA) yang dicadangkan dalam kajian ini menggunakan matriks

keahlian kabur dalam gugusan kabur untuk mencadangkan pendekatan baru bagi gugusan yang membenamkan penyelesaian pasca pengurangan heuristik untuk menangani kelemahan yang disebutkan di atas. Kaedah ini adalah EFCA yang cuba mencari gugusan bentuk yang serupa dengan ketepatan yang lebih baik. Ia memperkenalkan pendekatan untuk pengurangan semula yang diperiksa bukan sahaja terhadap diskriminasi neutron dan rayma tetapi juga pada pelbagai set data. Hasil daripada kaedah yang dicadangkan dinilai berdasarkan kaedah C-means kabur tradisional dan satu lagi kaedah gugusan yang terkenal, iaitu, K-means. Untuk diskriminasi neutron dan sinar gamma, EFCA meningkatkan ketepatan Rand index (RI) hampir 8%. Untuk kumpulan data multivariat dan siri masa lain yang digunakan dalam kajian ini, hasil menunjukkan pencapaian ketepatan yang ketara untuk sesetengah set data pasca pengurangan heuristik, mengakibatkan ketepatan RI 100% untuk sesetengahnya.

Kata kunci: Gugusan, Gugusan Kabur, Pembelajaran Tidak Terselia, Pra-pemprosesan, Pengurangan, diskriminasi Neutron dan sinar Gamma

ACKNOWLEDGMENTS

I would like to extend my special gratitude to my supervisor, Professor Dr. Teh Ying Wah, who supported, encouraged, and patiently guided me throughout this marvellous and memorable journey and contributed immensely to the works that have been presented in this thesis. I have been fortunate to have a supervisor who cared so much. I would also like to thank Dr. Saeed Aghabozorgi, who contributed as a co-supervisor in the first two years of my Ph.D.

Furthermore, I wish to thank my parents and my brothers, who have sacrificed and supported me during my studies. Special thanks go to my mother, Fatemeh Zahedifar, and my father, Seyed Mohammadreza Shirkhorshidi, who provided far more than the usual mental support. I have been fortunate enough to have them as my mentors and guides, who have contributed much to forming and shaping my research in various ways.

I also want to thank my wife, who has encouraged and supported me during this journey.

In addition, I would like to thank all the anonymous reviewers from the conferences and journals for their time and effort and for providing valuable comments that helped to improve the quality of the submitted research.

This thesis is dedicated to my parents for their unconditional love, endless support, and encouragement.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

ADCs	:	Analog to Digital Converters
ANN	:	Artificial Neural Networks
ANOVA	:	ANalysis OF VAriance
ARI	:	Adjusted Rand Index
BCFCM	:	Bias-Corrected Fuzzy C-Means
CCC	:	Cross-Correlation Clustering
CFCM	:	Credibility Fuzzy C-Means
CLIQUE	:	Clustering In QUEst
CSM	:	Cluster Similarity Measure
DB	:	Davies Bouldin
DBN	:	Dynamic Bayes Nets
DFC	:	Dynamic Fuzzy Cluster
DFT	:	Discrete Fourier Transform
DOFCM	:	Density Oriented Fuzzy C-Means
DTW	:	Dynamic Time Warping
DWT	:	Discrete Wavelet Transform
ECM	:	Evolving Fuzzy Clustering Method
ED	:	Euclidean Distance
EDR	:	Edit Distance on Real sequence
EFCA	:	Evolving Fuzzy Clustering Approach
ERP	:	Edit Distance with Real Penalty
FCM	:	Fuzzy C-Means
FCM-sigma	:	Robust Fuzzy C-means
FLICM	:	Fuzzy Local Information C- Means

FOM	:	Figure Of Merit
FPCM	:	Fuzzy-Possibilistic C-Means
FPGA	:	Field Programmable Gate Array
FS	:	Fuzzy Systems
GKA	:	Gustafson Kessel Algorithm
GMKIT2-FCM	:	Genetic-based improved Multiple Kernel Interval Type-2 FUzzy C-means clustering
GT2-FCM	:	General Type-2 Fuzzy C-Means
HMM	:	Hidden Markov Models
IFC	:	Incremental Fuzzy Clustering
IFCM	:	Intuitionistic Fuzzy C-Means
IFCM-sigma	:	Robust Intuitionistic Fuzzy C-means
IFE	:	Fuzzy Intuitionist Entropy
IFS	:	Intuitionist Fuzzy Set
IT2-FCM	:	Interval Type-2 Fuzzy C-Means
IVIFS	:	Interval Value Intuitive Fuzzy Set
kEFCM	:	kNN-Based Dynamic Evolving Fuzzy Clustering Method
KIFCM	:	Kernel Intuitionistic Fuzzy C-Means
kNN	:	k-Nearest Neighbors
KT2FCM	:	Kernel Type 2 Fuzzy set Fuzzy C-Means
KWFLICM	:	Kernel Weighted Fuzzy Local Information C- Means
LCSS	:	Longest Common Subsequence
MCM	:	Mixed C-Means
MDFCM	:	Multi-dimensional Fuzzy C-Means
MKFC	:	Multi Kernel Fuzzy Clustering
NC	:	Noise Clustering Algorithm

NIMs	:	Nuclear InstruMent units
NMI	:	Normalized Mutual Information
OFCM	:	Online Fuzzy C-Means
PAA	:	Piecewise Aggregate Approximation
PCM	:	Possibilistic Clustering Method
PFCM	:	Possibilistic Fuzzy C-Means
PGA	:	Pulse Gradient Analysis
PMT	:	PhotoMultiplier Tube
PSD	:	Pulse Shape Discrimination
RI	:	Rand Index
RMSSTD	:	Root-Mean-Square Standard Deviation
RNN	:	Recurrent Neural Networks
RP	:	Refinement Process
SAX	:	Symbolic Aggregate approXimation representation
SFCM	:	Spatial Fuzzy C-Means
SpADe	:	Spatial Assembling Distance
SpFCM	:	Single pass Fuzzy C-Means
STING	:	STatistical INformation Grid
T2FCM	:	Type 2 Fuzzy Sets Fuzzy C-Means
TOF	:	Time Of Flight
UFP-ONC	:	Unsupervised Fuzzy Partitioning-Optimum Cluster Number
WIPFCM	:	Weighted Image Patch-Based FCM

INTRODUCTION

1.1 Background: Radiation Detectors and Time Series Clustering

In nuclear physics, differentiating between neutron and gamma-ray pulses is a necessary procedure. The main task of radiation detectors is to discriminate and count neutron and gamma rays. These detectors are broadly used in different applications such as space research, mines, cultural heritage analysis, tomographical imaging, nuclear material control, international safeguarding, and national security (Amiri, Přenosil, Cvachovec, Matěj, & Mravec, 2014; Shan, Chu, Ling, Cai, & Jia, 2016; Uchida et al., 2014; Yousefi, Lucchese, & Aspinall, 2009).

The main problems in distinguishing neutrons from background gamma rays are as follows: first, signal pile-up and second, the difference in the energy of emitted signals that makes discrimination a challenging task. Traditional analogue pulse discrimination methods are less flexible and more time consuming than novel digital approaches. Digital technology has some significant privileges, such as clarity of energy, increased throughput, smaller size, easy upgrading and updating, automatic critical adjustments, multitasking operations, and automatic testing and verification.

After the recent emergence of digital methods of discrimination between neutrons and gamma rays, machine learning and artificial intelligence methods are also finding their way into this area. However, the machine learning methods used for discriminating between neutron and gamma rays are mostly traditional algorithms (Sanderson, Scott, Flaska, Polack, & Pozzi, 2012; Savran, Löher, Miklavec, & Vencelj, 2010; Yildiz & Akkoyun, 2013). These methods, as mentioned in the literature, and to the best of our knowledge, are used as suggested in computer science, and they are not improved to address specific needs in the area of discrimination. These methods consequently suffer from low accuracy. To address this gap, this thesis aims to increase the accuracy of

clustering approaches for the discrimination problem by taking into account the specific needs of this method.

Clustering is an unsupervised method of machine learning that has had a profound effect on various areas of knowledge discovery by addressing real-life problems (Aghabozorgi et al., 2014). The aim of clustering is to extract useful patterns among a series of data to prepare credible information; therefore, it can be applied as a base for the decision-making process (Liao, 2005).

Clustering can be defined as a method by which similar data are positioned in the same group, and different data are situated in different groups. The process of this method is unsupervised; that is, it does not need human supervision. Xu et al. (R. Xu & Wunsch, 2005) believe that there is no generally agreed explanation for clustering; however, they represent a short definition of the task of clustering algorithms: "Clustering algorithms divide data into a certain number of clusters (categories, groups or subsets) in which a cluster is described by considering the internal homogeneity and the external separation." This definition of a cluster is common amongst several other researchers in the field (D'Urso, De Giovanni, & Massari, 2015; Gower, 1971; Rao, 1971). Clustering has been used in a range of fields from software science (word analysis, picture segmentation, internet mining) and health sciences (microbiology, genetics, biology) to environmental sciences (geology, distant sensing), humanities, and sociology. Furthermore, clustering is used as a pre-processing stage for other data mining methods. The more precise definition of clustering is represented below.

Definition 1.1 Clustering: Given a dataset $D = \{v_1, v_2, \dots, v_n\}$ of data vectors and an integer value k, the clustering problem is to define a mapping $f: D \to \{1, \dots, k\}$, where each v_i is assigned to one cluster $C_j, 1 \leq j \leq k$. A cluster C_j contains precisely those data

vectors mapped to it; that is, $C_j = \{v_i \mid f(t_i) = C_{j,}\ 1 \leq i \leq n, and v_i \in D\}$. Moreover, v_1, v_2 represent two data vectors defined as follows:

$$v_1 = \{x_1, x_2, \dots, x_n\}$$

$$v_2 = \{y_1, y_2, \dots, y_n\}$$

where x_i, y_i are called attributes.

Definition 1.4 Fuzzy clustering: Suppose that $D = \{v_1, v_2, \dots, v_n\}$ is a set of data points. A fuzzy set A is formed if a function $fA: D \rightarrow [0,1]$ exists such that each element $a \in A$ is of the form $fA(t) = a$, for some $t \in T$. That is, each data point in T is assigned a value between 0 and 1, which describes its degree of membership or the probability of its placement in the set A. Fuzzy clustering, then, results in data objects belonging to one or more clusters and their membership in a particular cluster corresponding to some probability. The results of a fuzzy clustering can be represented by the $k \times n$ matrix U. The condition for fuzzy clustering is that for each entry of U, the form $u_{ji} \in \{0, 1\}$ exists, where $j \in \{1, \dots, k\}$ and $i \in \{1, \dots, n\}$ index the cluster and data point, respectively. Each row corresponds to clusters and each column to data points, and $\sum j = 1$.

Definition 1.5 Evolving fuzzy clustering of time series: In traditional clustering approaches, for a set of data points $D = \{v_1, v_2, \dots, v_n\}$, all of the data points would be clustered in one run. In contrast, in evolving clustering, data points are clustered gradually in a set of epochs $E = \{e_1, e_2, \dots, e_m\}$.

1.2 Motivation

Clustering time series data is challenging because they are often high-dimensional, and the datasets are enormous (Rani & Sikka, 2012). Neutron and gamma-ray pulses are also high-dimensional, and accurately discriminating them is the main problem. These pulses

are a set of energy levels of the particle that meets the scintillator; this set is collected over time and can be considered as a long time series. The pulse-shape discrimination (PSD) of n/gamma-ray has always been a challenging task, and distinguishing between high-energy and low-energy pulses requires different methods because of the difference in energy fluctuations (Amiri et al., 2014; Sosa, Flaska, & Pozzi, 2016). The clustering goal of the time series is to divide a large time series dataset into k clusters. Pulse-shape discrimination is a specific case of time series clustering when $k = 2$ because it wants to divide the dataset into two clusters of n/gamma-ray. As a result, time series clustering was used in some research works (Savran et al., 2010; Sayal & Kumar, 2011).

The goal of time series clustering is to detect those data that are similar to one another and then group them in clusters. The temporary order, large dimension, and outliers are significant problems associated with clustering time series data (Rani & Sikka, 2012). To cluster similar time series, a process of similarity matching needs to occur to calculate the similarity of the whole time series. This process is called whole time series clustering, in which the whole sequence of a time series is studied when the distance is calculated. However, calculating similarity measures is not a simple task, because of the noise, outliers, and shift, all of which make it a great challenge (Zakaria, Mueen, Keogh, & Young, 2016).

The clustering method coupled with a suitable similarity measure for dealing with the temporal order issue and dimensionality reduction that is supposed to tackle the last two issues. In similarity measure calculations, the time series length has a direct effect on the complexity of the computation. As a result, dimensionality reduction can significantly reduce the execution time for the clustering. On the other hand, when assessing the interval between two unprocessed time series variables when they carry noise, it is likely that the clusters will be similar in terms of noise instead of in their shape, consequently

has a dramatic influence on distorted clusters (Ratanamahatana, Keogh, Bagnall, & Lonardi, 2005). Choosing the appropriate dimensionality reduction method is thus vital, since it has an impact on final clusters, and it can potentially increase or decrease the computational complexity. By considering neutron and gamma-ray pulses as time series data, PSD shares the same set of challenges, and solving these problems motivated the researcher to conduct this study to improve neutron and gamma-ray discrimination.

1.3 Problem Statement

The problem statement for this study is as follows:

"Existing clustering algorithms have low accuracy and noise influences on the clustering outcome".

The low accuracy in pulse discrimination (time series clustering) was addressed to solve this problem. The accuracy of time series clustering generally suffers from unsuitable clustering algorithms, inaccurate distance measures, and inappropriate or untreated raw data.

Researchers have demonstrated that traditional machine learning algorithms produce an acceptable quality when they are used for particle discrimination (Akkoyun, 2013; Ronchi et al., 2009; Yu et al., 2015). However, much more work should be done for the discrimination problem and generally for time series because of their unique structure (Lin, Keogh, Lonardi, Lankford, & Nystrom, 2004; Savran et al., 2010).

1.4 Research Objectives

The primary objective of this research is to enhance the accuracy of clustering by introducing an evolving method that utilises the fuzzy membership matrix to divide the clustering task into multiple epochs. This main dataset that this research utilises is the neutron and gamma-ray dataset; however, the method has been evaluated on multiple

openly accessible datasets to prove its versatility. The objectives of this research are as follows:

1. to develop a new method for clustering that is more accurate for neutron and gamma-ray pulses;
2. to evaluate the capability of the suggested method for improving the accuracy of the clustering; and
3. to improve the performance of neutron and gamma-ray clustering (discrimination).

1.5 Research Questions

The following questions are addressed in this study to achieve the research objectives:

Objective 1: *To develop a new method for clustering that is more accurate for neutron and gamma-ray pulses.*

RQ1: How does one develop a clustering approach that yields a more accurate clustering result?

Objective 2: *To* evaluate *the capability of the suggested method for improving the accuracy of the clustering.*

RQ2: What is the influence of this method on the accuracy of continuous data (whether it is time series or multivariate)?

Objective 3: *To improve the performance of neutron and gamma-ray clustering (discrimination).*

RQ3: How can this method improve neutron and gamma-ray discrimination?

1.6 Scope of Research

To achieve the objective of the research in the designated period, the research scope must be clarified as follows:

1. The focus will be on whole time series clustering and the discrimination of neutron and gamma-ray pulses coming from the liquid base scintillator of type BC501A. The method should, however, apply to other types of detectors showing a particle-dependent pulse shape (Savran et al., 2010), as well as other datasets with continuous data.
2. Since accuracy in the discrimination problem plays a vital role, the target of this research is the improvement of the accuracy of the clusters.
3. This research focuses on finding the clusters based on their shapes; for this purpose, whole time series matching or "whole time series clustering" will be utilised (see Chapter 2).

1.7 Chapter Organisation

This section depicts the framework of the dissertation. The rest of the thesis is organised as follows:

Chapter 2 first provides some definitions of the major components of the study, such as clustering and fuzzy clustering. Then, the focus shifts to the researches and methods that are discussed in the fuzzy clustering area, and clustering time series and fuzzy time series clustering are addressed. This is followed by a review of existing studies on neutron and gamma-ray discrimination and its links with time series clustering.

Chapter 3 reviews the methodology employed in this study, clarifying how the objectives will be met and how the research questions will be answered. This chapter covers the implementation stages for the evolving fuzzy clustering approach (EFCA) for

neutron and gamma-ray discrimination. Furthermore, the evaluation plan is discussed to examine the quality of the proposed model.

Chapter 4 discusses the application of the proposed EFCA on various datasets along with its implementation for discriminating neutrons and gamma-rays. This chapter deals with the primary objective of this research, which proposes a more precise clustering algorithm than traditional clustering methods that is used for discrimination. The section also includes the experimental results, a discussion of the datasets used, and an evaluation of the suggested method along with the pre-processing of the information.

Chapter 5 concludes the study by reviewing how the research objectives were fulfilled and answering the research questions. The main contributions of the study are discussed, and possible future works are introduced.

CHAPTER 2:

LITERATURE REVIEW

2.1 Introduction

This study introduces an evolving approach using fuzzy clustering on time series data to solve the neuron and gamma-ray discrimination problem. Details about the method are discussed in Chapter 3; however, before elaborating on the new approach, this chapter reviews the research area of fuzzy clustering. First, definitions are provided for the major components of the study, such as clustering and fuzzy clustering; then, to contextualise the literature review, the focus shifts to the researches and methods that are discussed in the fuzzy clustering area. Time series, time series clustering, and fuzzy time series clustering are also discussed. Researchers have been working on fuzzy clustering by focusing on different perspectives. Some have focused on the type of time series, and some have focused on areas of application, while others have emphasise the method of fuzzy clustering itself. Since the application of the proposed method falls into the area of neutron and gamma-ray discrimination, clustering approaches that have been used in this area are also reviewed at the end of this chapter.

2.2 Clustering

Nowadays, the globe is filled with data. People find a great deal of information every day and store or describe it as data. A common way in which to cope with this information is to classify it or place it into a set of groups or clusters. This method is as ancient as the human need to classify and describe the outstanding traits of individuals and objects with a type, to identify groups of classes, and to assign items within them to suitable groups. Cluster identification is one of the essential methods for pattern recognition; in machine learning, this automatic process is called clustering. Most researchers define a cluster as internal similarity and external separation (Gordon, 1999; Jain & Dubes, 1988; Rao,

1971; Rui Xu & Wunsch, 2005). The primary characteristic of these clusters is that the objects in the same clusters are interrelated, while the objects in distinct clusters are distinct (Rodríguez-Fernández, Menéndez, & Camacho, 2017). Data points must be divided into clusters depending on their resemblance in unlabelled datasets. The data are structured in this manner into an effective representation that symbolises the sampled population. The following steps typically provide a clustering operation model (Jain & Dubes, 1988):

- depiction of patterns (preferably with identification and choice of features);
- definition of model proximity criteria suitable for the field;
- aggregation or pairing;
- integration of information (if necessary), and
- evaluation of performance (if necessary).

Clustering is widely used in pattern recognition in various fields (Shirkhorshidi, Aghabozorgi, Wah, & Herawan, 2014) including categorising, problem-solving and artificial intelligence, data retrieval (Khalifi, Cherif, Qadi, & Ghanou, 2019; Ye, Luo, Dong, He, & Min, 2019), image processing (Chen, Sun, Palade, Shi, & Liu, 2019), and pattern identification (Abd-Elaal & Hefny, 2013; Jain & Dubes, 1988; Mecca, Raunich, & Pappalardo, 2007; Nie & Zhang, 2013). It has a rich background in other fields as well, from scientific applications and engineering to advertising sciences.

Definition: Clustering

Given a dataset $D = \{d_1, d_2, \dots, d_n\}$ of data instances and an integer value k, the clustering problem is to define a mapping $f: D \to \{1, \dots, k\}$, where each d_i is assigned to one cluster $C_j, 1 \leq j \leq k$. A cluster C_j contains precisely those data instances mapped to it; that is, $C_j = \{v_i \mid f(t_i) = C_j, 1 \leq i \leq n, and v_i \in D\}$. The clustering structure is officially

considered to be a set of subsets $C = \{C_1, C_2, \dots, C_k\}$ of D. Any instance in D accordingly applies to just one sub-set. However, for all these algorithms, the number c of clusters must be pre-assumed. Moreover, the number k should generally be unknown; therefore, the method of finding the optimal k is highly important. This type of issue is usually referred to as the validity of the cluster.

Machine learning in information extraction is a sub-category divided into unsupervised and supervised learning methods, each of which is designed to serve distinct aims. While supervised machine learning is predictive, unsupervised methods are descriptive. Clustering would be an unsupervised learning approach, while classifications are supervised.

In the literature, clustering is associated with various concepts, namely, unsupervised training (Guan, Yuen, & Coenen, 2019), vector quantisation (Shastri et al., 2019), and numerical taxonomy (Aparicio-Ruiz, Martín, Martín, & Achedad, 2019). The significance and collaborative character of clustering is evident throughout its vast literary works.

2.3 Clustering Methods

The primary reason that distinct clustering methods exist is the absence of an accurate definition of the concept of a "cluster". Over the years, different perspectives have inspired different clustering algorithm taxonomies (Aghabozorgi, Shirkhorshidi, Ying Wah, Seyed Shirkhorshidi, & Ying Wah, 2015; Berkhin, 2006; Everitt, Landau, & Leese, 2011; A. K. Jain, Murty, & Flynn, 1999). As a result, several clustering methods have been developed, each of which has a distinct preparation theory. Researching the literary works provides many solutions to clustering (Jain & Dubes, 1988; Jain, Murty, & Flynn, 1999; Kolatch, 2001; Rao, 1971; Rui Xu & Wunsch, 2005); numerous clustering studies exist, such as the following: (Daxin Jiang, Chun Tang, & Aidong Zhang, 2004), (Xu &

Wunsch, 2005), and (Hruschka, Campello, Freitas, & de Carvalho, 2009a). In 1998, Farley and Raftery divided clustering methods into two significant categories, namely, hierarchical and partitional, and more recently, Han and Kamber (2006) expanded this categorisation to include three more categories: density-based, model-based, and grid-based clustering methods. Figure 2 presents a summary of clustering methods, and the various methods of clustering are elaborated in the rest of this section.

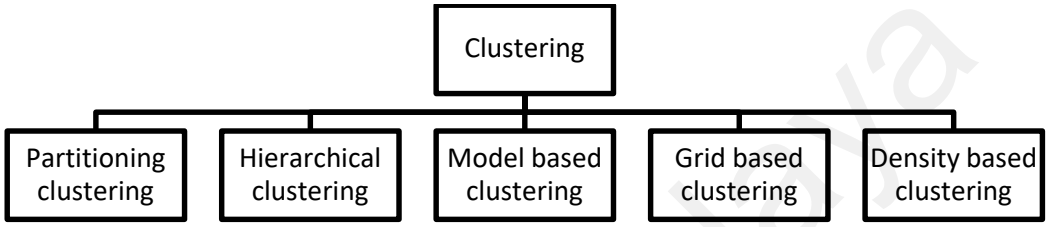

Figure 2: A summary of clustering methods

2.3.1 Hierarchical clustering

The hierarchical clustering method has a strong visualisation relative to other methods of clustering (Keogh & Kasetty, 2003). It generates a nested hierarchy of associated classes of items considering the pairwise distance matrix of the objects. The strength of this approach is that the user does not need to introduce parameters , such as the number of the groups, in advance. However, a shortcoming of this method is that its implementation is restricted to small datasets because of its quadratic computing complexity (Lin, Keogh, & Truppel, 2003). Hierarchical methods are categorised into two distinct kinds, namely, agglomerative and divisive methods, where clusters are formed by repeating the partitioning of a subject by way of top-down (divisive) or bottom-up (agglomerative) methods. On the one hand, agglomerative hierarchical clustering has a bottom-up structure; therefore, each object represents its cluster in the beginning. Then, clusters are continually combined until the cluster's required larger structure is created. On the other hand, the design of the divisive method is the reverse; that is, a top-down system is applied. All items initially

correspond to one cluster in divisive hierarchical clustering. Then, the cluster splits into smaller sub-clusters that break into their sub-category. This method proceeds until the structure of the required cluster is formed. Figure 2-1 demonstrates the divisions of hierarchical clustering.

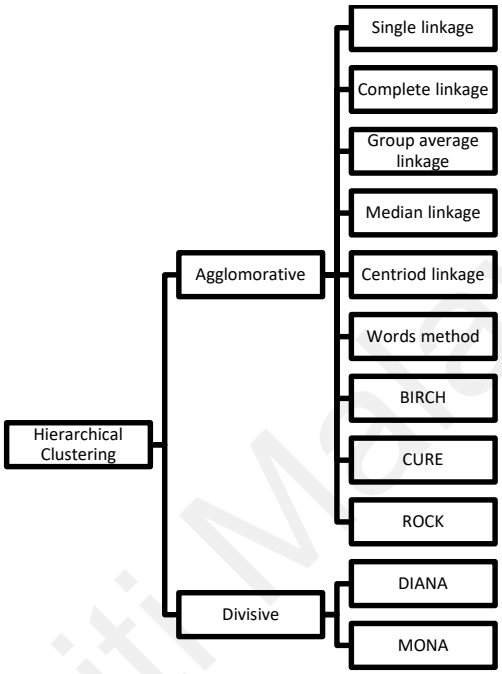

Figure 2-1: Divisions of hierarchical clustering

2.3.2 Density-based clustering

Density-based methods (such as DBSCAN) continue to create a group until the cluster density (number of items or information points) reaches a certain limit. The items belonging to each group are presumed to be selected from a precise distribution of probability (Banfield & Raftery, 1993). A mixture of various distributions ought to be the general distribution of data. These methods attempt to define the clusters and their data distribution.

2.3.3 Model-based clustering

Model-based clustering suggests the use of (finite) clustering systems for clustering performance and attempts to improve the match between the data provided and some

mathematical designs. This method is different from conventional clustering that classifies data in that it provides each cluster's feature characteristics, which represent a category or class. The method is intended to model an unknown distribution (or cluster) as a mixture of simpler distributions, sometimes called basis distributions. Decision trees and neural networks seem to be the most commonly utilised model-based methods.

2.3.4 Grid-based method

The grid-based method identifies a set of grid cells, divides objects into a fixed number of grid units, calculates each cell's density, and then removes cells with a lower predefined threshold density. Moreover, it forms clusters from nearby groups of dense cells (usually minimising a specified objective function). The approach's main advantage is its quick processing time (Han et al., 2006). However, a clustering difficulty relates to the number of grid cells that are occupied and not the number of items in the dataset. Some popular clustering methods based on the grid are the STatistical INformation Grid (STING) approach and the Clustering In QUEst (CLIQUE) approach (Saini & Rani, 2017).

2.3.5 Partitioning clustering

Partitioning methods which are known as partitional methods are some of the data clustering methods that are most commonly used. They involve the movement of information from one group to another, beginning with an original partitioning**.** In these methods of clustering, data objects are typically directly assigned to a pre-set number of clusters (C partitions) that the user should pre-set, although selecting the number of required clusters is a problem. Partitional methods usually produce clusters based on some similarity measured by the objective function.

A comprehensive iterator procedure of all feasible partitions is required to attain universal optimality in partitional clustering. However, as this is not feasible, in the

manner of iterative development, a highly specific method is used. A relocation technique, for example, migrates sample objects iteratively between *k* clusters. Partitional clustering itself is divided into two types of clustering, namely, crisp (hard) and fuzzy (soft). In crisp clustering, data points correspond to just one cluster at the moment; therefore, the clusters are separated into a hard clustering. In contrast, fuzzy clustering broadens the concept, where each piece of data is correlated with each cluster by way of the membership feature (Zadeh, 1965). In fuzzy clustering, data points can be members of several clusters, each to a different degree (A. K. Jain, Murty, Fynn, et al., 1999). Figure 2-2 illustrates different partition methods for clustering.

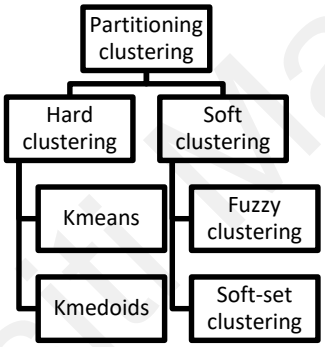

Figure 2-2: Partitioning clustering methods

The focus of this study is on partitional clustering – specifically its fuzzy approaches. In the following sections, different components of partitional clustering are discussed, as illustrated in Figure 2-3.

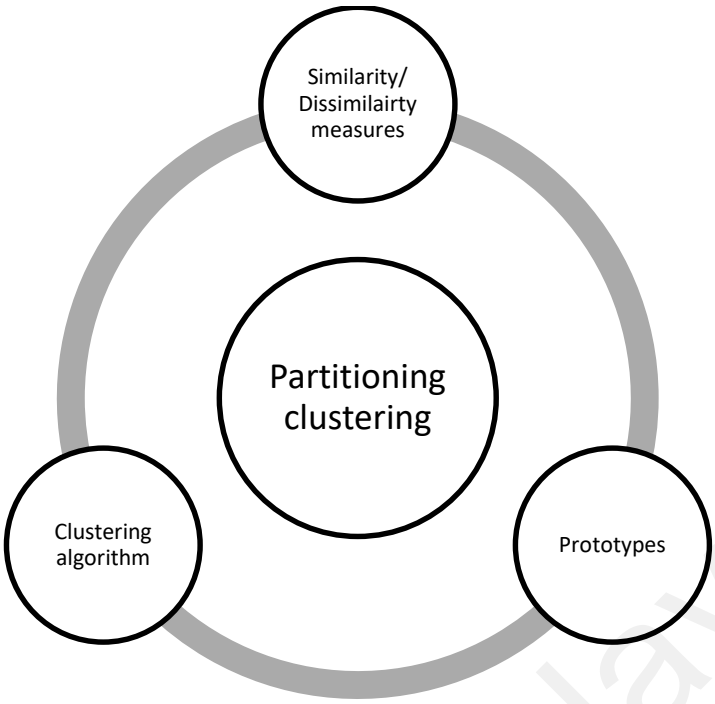

Figure 2-3: Different component of partitional clustering

2.3.5.1 Similarity or dissimilarity measure

Clustering is a collection of comparable objects or items; therefore, it may be necessary to determine whether two items are similar or different. One major consideration is the type of method that should be utilised to specify similarity or dissimilarity between a pair of objects, or between an object and a cluster, or between two clusters. To assess this connection, two primary measures are distance measurements and similarity measurements.

(a) ***Dissimilarity (distance) measures:***

Many clustering methods make use of distance measurements for defining the resemblance or difference between any pair of objects. The distance between x_i and x_j can be represented as $d(x_i, x_j)$. A legitimate distance measurement should be symmetrical and achieve its minimum value in the event of exactly similar points (usually zero) (Rokch & Maion, 2005).

(b) ***Similarity measures:***

The similarity function $s(x_i, x_j)$ compares the two x_i and x_j vectors and is a concept contrary to that of distance. This function should be symmetrical – that is, $s(x_i, , x_j) = s(x_j, x_i)$ – it should also have a high value if x_i and x_j are "similar", and it should yield the highest value for identical vectors (R. Xu & Wunsch, 2005). The following table presents the various similarities and dissimilarities, their applications, and their advantage and disadvantages for continuous data.

Table 2-1: Similarity and dissimilarity measures for continuous data (in time complexity; n is the number of dimensions of x and y)

Distance Measure	Equation	Time complexity	Advantages	Disadvantages	Applications
Euclidean Distance	$d_{euc} = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{\frac{1}{2}}$	$O(n)$	It is common and easy to compute, and it works well with datasets with compact or isolated clusters (Gan, Ma, & Wu, 2007; A. K. Jain, Murty, & Flynn, 1999).	It is sensitive to the outliers (Gan et al., 2007; A. K. Jain, Murty, & Flynn, 1999).	*K*-means algorithm, fuzzy *C*-means algorithm (Ji, Xie, & Ping, 2013).
Average Distance	$d_{ave} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2\right)^{\frac{1}{2}}$	$O(n)$	It is better at handling outliers in comparison with Euclidean distance (Legendre & Legendre, 2012).	Variables contribute independently to the measure of distance. Redundant values could dominate the similarity between data points (Hand, Mannila, & Smyth, 2001)	*K*-means algorithm
Weighted Euclidean	$d_{we} = \left(\sum_{i=1}^{n} w_i(x_i - y_i)^2\right)^{\frac{1}{2}}$	$O(n)$	The weight matrix allows one to increase the effect of more important data points in comparison to less important points (Hand et al., 2001).	Same as the disadvantages of average distance.	Fuzzy *C*-means algorithm (Ji et al., 2013)
Chord	$d_{chord} = \left(2 - 2\frac{\sum_{i=1}^{n} x_i y_i}{\|x\|_2 \|y\|_2}\right)^{\frac{1}{2}}$	$O(3n)$	It can work with un-normalised data (Gan et al., 2007).	It is not invariant to linear transformation (R. Xu & Wunsch, 2005).	Ecological resemblance detection (Legendre & Legendre, 2012)
Mahalanobis	$\mathrm{d_{mah}} = \sqrt{(\mathrm{x}-\mathrm{y})\mathrm{S}^{-1}(\mathrm{x}-\mathrm{y})^{\mathrm{T}}}$	$O(3n)$	Mahalanobis is a data-driven measure and can lighten the distance distortion caused by a linear combination of attributes (Legendre & Legendre, 2012) .	It can be expensive in terms of computation (R. Xu & Wunsch, 2005)	Hyper ellipsoidal clustering algorithm (Mao & Jain, 1996)
Cosine Measure	$\mathrm{Cosine}(\mathrm{x},\mathrm{y}) = \frac{\sum_{i=1}^{n} \mathrm{x_i y_i}}{\|\mathrm{x}\|_2 \|\mathrm{y}\|_2}$	$O(3n)$	It is independent of vector length and invariant to rotation (R. Xu & Wunsch, 2005).	It is not invariant to linear transformation (R. Xu & Wunsch, 2005).	Mostly used in document similarity applications (Han et al., 2006; R. Xu & Wunsch, 2005).
Manhattan	$d_{man} = \sum_{i=1}^{n}(x_i - y_i)$	$O(n)$	It is commonly used and, similarly to other Minkowski-driven distances, it works well with datasets with compact or isolated clusters(Gan et al., 2007).	It is sensitive to the outliers (Gan et al., 2007; A. K. Jain, Murty, & Flynn, 1999)	*K*-means algorithm
Mean Character Difference	$d_{MCD} = \frac{1}{n}\sum_{i=1}^{n}\|x_i - y_i\|$	$O(n)$	*It results in accurate outcomes using the *K*-medoids algorithm.	*Low accuracy for high-dimensional datasets using *K*-means.	Partitioning and hierarchical clustering algorithms.
Index of Association	$d_{IOA} = \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{x_i}{\sum_{i=1}^{n} x_i} - \frac{y_i}{\sum_{i=1}^{n} y_i}\right\|$	$O(3n)$	-	*Low accuracy using *K*-means and *K*-medoids algorithms.	Partitioning and hierarchical clustering algorithms
Canberra Metric	$d_{canb} = \sum_{i=1}^{n}\frac{\|x_i - y_i\|}{(x_i + y_i)}$	$O(n)$	*It results in accurate outcomes for high-dimensional datasets using the *K*-medoids algorithm.	-	Partitioning and hierarchical clustering algorithms
Czekanowski Coefficient	$d_{czekan} = 1 - \frac{2\sum_{i=1}^{n} min(x_i, y_i)}{\sum_{i=1}^{n}(x_i + y_i)}$	$O(2n)$	*It results in accurate outcomes for medium-dimensional datasets using the *K*-means algorithm.	-	Partitioning and hierarchical clustering algorithms
Coefficient of Divergence	$d_{canb} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - y_i}{x_i + y_i}\right)^2\right)^{\frac{1}{2}}$	$O(n)$	*It results in accurate outcomes using the *K*-means algorithm.	-	Partitioning and hierarchical clustering algorithms

2.3.5.2 Evaluation Measures

One of the problems after clustering is to assess whether a particular clustering is accurate. Bonner (1964) was one of the first to claim that there was no widespread description of what would be a perfect clustering. The evaluation is primarily a matter of

opinion; nevertheless, the literature has recognised some evaluation criteria. The evaluation methods are usually split into three types: internal indexes, external indexes, and relative metrics. A similarity examination tests and estimates the relative value of two pieces of data. In (A. Jain & Dubes, 1988) and (C. H. Chen, Pau, Wang, & Dubes, 1993), the factors used for assessment are addressed in detail. In the next two sections, internal and external quality indices are investigated.

(a) ***Internal Quality Criteria***

The outcomes of clustering algorithms are generally assessed using internal validation measurements. An internal validity measurement uses the results of the clustering itself to measure the results' validity. It attempts to determine whether the structure is essentially suitable for data, often based on two criteria: **compactness** and **separation**. Internal evaluations generally apply some indices of resemblance to evaluate cluster density. This method usually evaluates intra-cluster similarity and inter-cluster distance, or it can use a combination of both, and neither of these requires any external source other than its dataset (Liu, Li, Xiong, Gao, & Wu, 2010; Qamar, 2014). On the one hand, compactness measures intra-cluster similarity. Intra-cluster compactness is calculated by using the variance of cluster or distance between objects in the cluster. On the other hand, a maximum value is expected to be the distance between different clusters measured by separation criteria. The distance measurement can be the centre-to-centre cluster distance or the distance between cluster data points. Some of the measures for internal validation are as follows (Sivarathri & Govardhan, 2014):

- the sum of squared error (SSE);
- other minimum variance criteria;
- scatter criteria;
- Condorcet's criterion;

- the *C*-criterion;
- the category utility metric; and
- edge cut metrics.

(b) ***External Quality Criteria***

An external validity measurement may be useful to examine how similar the structures of the clusters (returned by the clustering algorithm) are to some predefined classification of the instances. It compares the obtained structure with the original one. Those external validity indicators are as follows (Rokch & Maion, 2005):

- mutual information based measure;
- precision-recall measure;
- the Rand index (RI); and
- the adjusted Rand index (ARI).

In this research, the RI is selected (Rand, 1971) as an external measure of validity. The RI is a common criterion for calculating the resemblance of the clustering model C1 with the clustering framework C2. The RI can also be considered as a percentage of the algorithm's correct decisions. Let *a* (true positive) be the number of sets of items in a certain C1 group and the certain C2 group, and let *b* (false positive) be the number of sets of items in the C1 group but not in the C2 group union with the number of sets of items within the identical group in C2, while they are not in the similar group in C1. *d* (true negative) is the number of sets of items in identical cluster on C2 but just not in same class in C1 or the number of pairs of items allocated to distinct groups of C1 and C2. It is possible to interpret *a* and *d* as similarities, and *b* and *c* as differences. The RI will be described as follows:

$$RAND = \frac{a+d}{a+b+c+d} \quad 2\text{-}1$$

or, in other words,

$$Rand\ Index\ = \frac{TP+TN}{TP+FP+FN+TN} \quad 2\text{-}2$$

The RI ranges from 0 to 1, and a high RI value indicates the highest accuracy. The RI will be 1 when there is a perfect agreement between the two partitions. However, a problem with the RI is that its estimated value of two random groups does not yield a consistent number (such as 0) when the two clusters do not completely match. To solve such a limitation, Hubert and Arabie (1985a) suggest the ARI.

ADJUSTED RAND INDEX

An issue with the RI matrix is that between the two random groups, the anticipated RI score is not a constant. Therefore, Hubert & Arabie (1985) recommend an adapted RI that aims to fix this issue by presuming a generalised hyper-geometric distribution as a random model. It also performs better than the RI and several other indicators (Milligan & Cooper, 1986; Steinley, 2004). The modified RI matrix has the highest value of 1 for random groups, and the predicted value is 0. The higher the ARI, the greater the agreement between the two partitions. Furthermore, for measuring agreement, the ARI is suggested, even if there are distinct cluster numbers for the comparative clusters. This approach has been satisfactorily used in the gene regulation database (Yeung, Fraley, Murua, Raftery, & Ruzzo, 2001; K. K. Yeung, Haynor, & Ruzzo, 2001). The ARI is calculated by the following equation:

$$ARI(C,G) = \frac{R\ -\ E[RI]}{max(RI)\ -\ E[RI]} \quad 2\text{-}3$$

2.4 Fuzzy Clustering

Conventional clustering approaches, also known as hard clustering approaches, group information and make partitions with differentiated limits; each object is required to belong only to one cluster. Clusters are consequently mutually exclusive in a hard clustering. On the other hand, fuzzy clustering is non-linear in principle, and each object can have a degree of membership to other clusters, which provides more natural partitioning approaches; this implies more decision-making alternatives for the fuzzy clustering approach (Li & Lewis, 2016). This notion expands and suggests more diverse results that are soft representations of clusters. The word "fuzzy" refers to the cluster overlap where each component in a primary source belongs, to an extent, to one or even more clusters. In this situation, utilising the membership matrix, each element is bound to all clusters, suggesting that each group is a fuzzy cluster of all components. m_{ij}, which represents membership in fuzzy clustering, issued to detect knowledge about item relationships and associated groups. There would be higher assurance in distributing the model to the cluster with higher membership scores. A hard clustering can also be attained through a threshold level of membership value from a fuzzy clustering (Rodríguez-Fernández et al., 2017).

Two main types of membership are generally considered in the fuzzy clustering literature: first, a relative type, called probabilistic membership, which indicates the percentage to which each cluster should be ascribed for a specified point, and second, an absolute or possible type, specifying the strength of the allocation to any group independent of the remainder (Masulli & Rovetta, 2006). The majority of analytical fuzzy clustering methods are derived from the main fuzzy *C*-means (FCM) algorithm. The FCM uses the probabilistic restriction that data point membership across classes sums to 1.

The Euclidean distance (ED) that is used in FCM, is widely used in various clusters to accomplish distance between elements in clusters and cluster centres, known as fuzzy membership (Abdulla & Al-Nassiri, 2015). It is the first choice for many applications because the ED results in spherically patterned clustering. Furthermore, Gustafson-Kessel has used the Mahalanobis distance to assess various cluster patterns such as the ellipsoidal group, and Gath and Geva subsequently introduced maximum likelihood measurements to evaluate a comparative ellipsoidal pattern chance (Gath & Geva, 1989; J. Li & Lewis, 2016). Researchers have created fuzzy *c* variants, the optimised fuzzy clustering method, the FCM algorithm, the Gustafson-Kessel method, and the Gath-Geva method, motivated by yielding results with a more sensible pattern compare to previous methods (Jantzen, Norup, Dounias, & Bjerregaard, 2005).

2.4.1 Types of fuzzy sets

In areas such as monitoring and rule-based logic, fuzzy set techniques are common, since they have the potential to depict undefined categories and notions naturally. The illustration of such undefined categories or notions is accomplished through membership features described in the relevant discourse context in Zadeh's formulation of the fuzzy set theory (Zadeh, 1965, 1978). In fuzzy clustering, subjects share a percentage of membership in multiple groups. Fuzzy clustering enables the step-by-step calculation of the membership of components in a cluster formed by the membership parameter between the interval of [0,1], and the total membership values for the study group will be 1. So, each element in fuzzy sets is assigned to $[0,1]$ through the membership function $\mu_A: X \rightarrow [0,1]$ in which $[0,1]$ means the real numbers from 0 to 1 (including 0 and 1). As a result, functions assigning membership values to elements are specified as either a Type-I, Type-II, or intuitionist fuzzy set (IFS) (Gosain & Dahiya, 2016). The following figure presents different fuzzy set theories.

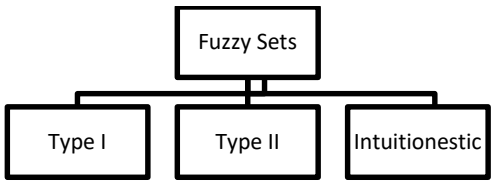

Figure 2-4: Different fuzzy set theories

2.4.1.1 Type-I fuzzy set

Datasets in real application can have many uncertain factors, which result in uncertainties in the fuzzy set membership functions. Type-I fuzzy sets utilise crisp membership functions that are not able to model such uncertainties. Moreover, in Type-I fuzzy sets, a specialist would dictate the degree of membership. As an example, in Type-I fuzzy sets, if one has three different green balls, then the first is green by a membership value of 65%, the second is green by 80%, and the third is green by 90%.

2.4.1.2 Type-II fuzzy set

Fuzzy set Type-II membership values themselves are defined by a fuzzy set. Zadeh (Zadeh, 1975) has presented this concept, which generalises Type-I fuzzy sets, thus enabling one to include uncertainty about the membership function in the fuzzy set approach (H. B. B. Mitchell, 2005; Yao & Weng, 2012). Fuzzy set Type-II was used to manage uncertainties in different contexts, where Type-I fuzzy sets are not performing satisfactorily. For example, Mitchell (H. B. Mitchell, 2003; H. B. B. Mitchell, 2005); Nakhostin (2012); and Zeng, Xie, and Liu (Zeng, Xie, & Liu, 2008) have applied the Type-II fuzzy set to manage pattern identification instability. Referring to the previous example, researchers cannot precisely determine the degree to which the characteristics are achieved in a Type-II fuzzy set. If one has three different green balls, for example, the first is green by 65% to 70%, the second is green by 80% to 95%, and the third is red by 90% to 95%. Type-II fuzzy set thus provides an interval fuzzy set.

The benefit of fuzzy sets and mechanisms in Type-II fuzzy sets is a three-dimensional membership feature to handle more uncertainty in actual problems (Mendel, 2007). Increased dynamism in a description implies enhanced capacity to logically and correctly manage incorrect data. A higher version of fuzzy association (e.g., Type-II) was conceived as one manner of increasing a relationship's fuzziness. One can extend this notion to Type-*n* of a fuzzy set (Qamar, 2014).

In the case of linguistics, for example, Type-II fuzzy sets allow one to address language uncertainties that can be stated as, "for every person, each word can imply different things".

2.4.1.3 Intuitionistic fuzzy set

Atanassov (K. T. Atanassov, 1986) extended Zadeh's fuzzy set by using two concepts to evaluate the component's membership and non-membership values, which pertain to the distance of [0,1], and its sum also relates to the same interval. On the other hand, it has been demonstrated that an IFS is more effective than a fuzzy set in coping with fuzziness and instability. However, in real life, it might not be true that the value of non-membership of an object in a fuzzy set is equivalent to 1 minus the value of membership, because some level of uncertainty may be present. Therefore, an IFS combines the mentioned level of uncertainty (and is described as one minus the total sum of the degree of membership and non-membership).

Many experts have investigated the IFS concept over the past decades, and they used it in different areas. Atanassov and Gargov (K. Atanassov & Gargov, 1989) added a general intuitive fuzzy set in standard fuzzy sets on the basis of a distance value and then proposed an intuitive fuzzy set according to the interval value (IVIFS). Then, in (Atanassov, 1986b; De, Biswas, & Roy, 2000; Deschrijver, Cornelis, & Kerre, 2004; Deschrijver & Kerre, 2007; Xu & Yager, 2006; Zeshui Xu, 2007), continuous studies

have been conducted regarding relationships, procedures, and operators linked to an IFS and IVIFS . Deschrijver and Kerre (Deschrijver & Kerre, 2003) have established relationships between I-Fuzzy sets, the interval-valued IFS, and L - FS (Goguen, 1967) and an IFS. A further fuzzy set named vague set similar to an IFS has been launched at (Bustince & Burillo, 1996; Gau & Buehrer, 1993). Furthermore, some techniques for measuring the correlation variables of an IFS and IVIFS were put in place (Bustince & Burillo, 1995, 1996b; Gerstenkorn & Mańko, 1991; Hong & Hwang, 1995; HUNG, 2003; Hung & Wu, 2002; Mitchell, 2004; Xu, 2006b). Xu (Xu, 2006c), in a detailed study of the correlation analysis of an IFS, stated that many of the current correlation methods could not guarantee that the IFS or IVIFS correlation ratio is equal to 1 only if these two are the same (Xu, Chen, & Wu, 2008).

2.4.2 Application of fuzzy clustering

Fuzzy clustering algorithms have been studied and applied in many different areas. They also turn into major cluster analysis methods. These applications open the door to research in fuzzy clustering. The literature recognises the practical importance of clustering in various disciplines such as classification, medicine, geography, finance, engineering technologies, and graphics processing. Fuzzy clustering has now been extensively researched and implemented in distinct scientific fields; for example, Type-II fuzzy sets and IFSs have been exploited in decision-making processes, the design of fuzzy relationship formulas, questioner processing, time series estimation, approximation operations, the equalisation of time variables, and portable robots regulation, amongst other things (Yao & Weng, 2012). An IFS has also been applied in various fields, such as decision-making processes (Castillo & Atanassov, 2019; Herrera, Martínez, & Sánchez, 2005; Hong & Choi, 2000; Li, Olson, & Qin, 2007; Liu & Wang, 2007; Pankowska & Wygralak, 2006; Saadati & Park, 2006; Szmidt & Kacprzyk, 2002, 2003; Xu, 2007c, 2007d, 2007b, 2006a; Xu & Yager, 2006), economics and society

studies (Meier, Pedrycz, & Portmann, 2019), medical diagnoses (Hu, Pan, Yang, & Chen, 2019; Szmidt & Kacprzyk, 2004), pattern recognition (Hung & Yang, 2007; Jin & Bai, 2019; Mitchell, 2005c; Wang & Xin, 2005; Xu et al., 2008), and robotic systems (Narayanamoorthy, Geetha, Rakkiyappan, & Joo, 2019). These methods have become the key tools for analysing a cluster.

2.5 Fuzzy Clustering Algorithm

From the last century, when fuzzy clustering was introduced, it has been applied in many disciplines. However, today, with the significant amounts of internet information transfer, improving fuzzy clustering methods is still an open issue.

Assume a set of n objects $X = \{x_1, x_2, \ldots, x_n\}$, where x_i is a d-dimensional point. A fuzzy clustering is a collection of k clusters, $\{C_1, C_2, \ldots, C_k\}$, and a partition matrix $W = \{w_{i,j} \in [0,1], for\ i = 1 \ldots n\ and\ j = 1 \ldots k\}$, where each element $w_{i,j}$ is a weight that indicates the value of membership of item i in the cluster of C_j.

Kaufman (1990) describes the fuzzy algorithm, which aims to minimise the following objective function, *J*, consisting of cluster memberships and distances.

$$J = \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} m_{i,k}^2 m_{j,k}^2 d_{i,j}}{2 \sum_{j=1}^{N} m_{j,k}^2} \qquad 2\text{-}4$$

where $m_{i,k}$ is the unknown membership of the object i in cluster k, and $d_{i,j}$ is the dissimilarity between objects i and j. Memberships are subject to limitations that they must all be non-negative and that memberships must be summed up to 1 for a single item. That is, memberships have the same limitations they would have if they were the probabilities that an item belongs to each group (and can be defined as such).

A wide range of proposed algorithms exist for fuzzy clustering; these methods and their respective fuzzy category are presented in Figure 2-5. Several variations and generalisations of the fuzzy clustering algorithm are discussed in the following sections.

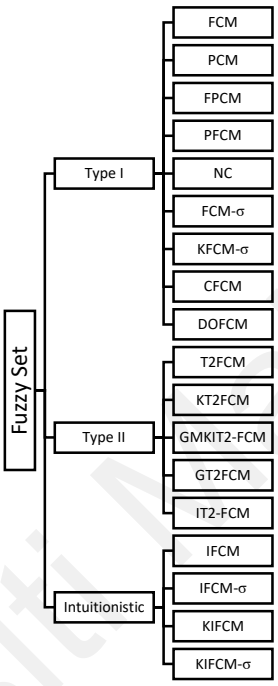

Figure 2-5: Fuzzy clustering method

2.5.1 Fuzzy *C*-means

It could be claimed that the most common fuzzy clustering method is the FCM. In 1984, Bezdek and his colleagues (Bezdek, Ehrlich, & Full, 1984) proposed the FCM as an expansion of Zadeh's (Zadeh, 1965) fuzzy set to resolve imprecision and uncertainty. It is a frequently used and effective method of clustering and classification. Fuzzy *C*-means is a fuzzy variant of *K*-means and is better at avoiding local minima than the hard *K*-means approach; as with *K*-means, the FCM also seeks to reduce the SSE. FCM utilises the ED to evaluate differences like *K*-means; this enforces a spherical cluster shell without considering the real data distribution. The most significant issue with fuzzy clustering is the membership feature design; it can be based on the decomposition of similarities or

cluster centroids. Furthermore, it uses fuzzy logic and fuzzy set theory concepts. In the FCM, an object can correspond to various clusters at the same time, with a certain degree known as the membership value, depending on the cluster centre's likelihood. Therefore, the membership value is an indicator of similarity. The membership degree for an item can be a value between 0 and 1; greater similarity results in a higher membership value (Choudhry & Kapoor, 2016). Traditional set theory can thus be viewed as a special case in which membership values are limited to either 0 or 1. A noticeable reality about this type of method is the defect within the underlying accepted model: each point in *X* is categorically grouped with other members of "its" cluster and hence bears no obvious similarity to other members of *X*. For this discussion, it suffices to notice that hard partitions of *Y* are special types of fuzzy partitions, whereby each data point is grouped without ambiguity with its intra cluster neighbours. Anomalies (noise or otherwise) usually make a group of "unclassifiable" points; however, most standard models do not have any natural mechanism to handle the effect of noisy data.

A fuzzy *c*-partition of *X* is one that defines the membership of each sample point through a membership function in all clusters that ranges from 0 to 1. Furthermore, the sum of the memberships for each sample point should be 1.

Let $Y = \{Y_1, Y_2 \ldots . Y_n\}$ be a sample of n observations in R^n (n-dimensional Euclidean space); Y_l is the l-th feature vector, and Y_{lj} is the j-th feature of Y_l. If c is an integer, $2 \leq c$, then a conventional (or "hard") c-partition of Y is a c-tuple $(C_1, C_2 , \ldots . , C_c)$ of subsets of Y that satisfies three conditions:

$$\mathrm{C_i} \neq \emptyset, 1 \leq i \leq c; \ (1a)$$

$$\mathrm{C}_i \cap \mathrm{C}_j = \emptyset; \ i \neq j \ (1b) \qquad 2\text{-}5$$

$$\cup_{i=1}^{c} \mathrm{C}_i = Y \ (1c)$$

Let U be a real $c \times n$ matrix, $U = [u_{il}]$. Here, U is the matrix representation of the partition $\{C_i\}$ in equation (1) in the situation

$$u_i(y_l) = u_{i,l} \begin{cases} 1;\ y_l \in Y_i \\ 0;\ otherwise \end{cases} \quad (2a)$$

$$\sum_{l=1}^{n} u_{i,l} > 0 \quad for\ all\ i; \quad (2b)$$

$$\sum_{i=1}^{c} u_{i,l} = 1 \quad for\ all\ l; \quad (2c)$$

2-6

Defuzzification will be utilised at the final stage of the clustering procedure to determine the groups. The FCM algorithm is repetitious, and to accomplish it, the centre of the cluster and the membership degree are frequently updated. By setting the cost equation, these upgrading equations are created. Have $X = \{x_1, x_2, x_3, \ldots, x_N\}$ represent a dataset with N items. It must be split into *c*-clusters by minimising the price function (Choudhry & Kapoor, 2016).

$$j = \sum_{j=1}^{N} \sum_{i=1}^{c} u_{i,j}^{m} \left\| x_j - v_i \right\|^2 \quad 2\text{-}7$$

where u_{ij} denotes the affiliation of x_j in the i^{th} cluster, v_i would be the ith centre of the cluster, $\|.\|$ is a norm metric, and "m" would be a constant. Moreover, parameter m determines the resulting partition's fuzziness. By attempting to take the derivative of the formula and making it equal to 0 through the use of the Lagrange technique, the following results are accomplished (Choudhry & Kapoor, 2016):

$$v_i \frac{\sum_{j=1}^{N} u_{ij}^{m} x_l}{\sum_{j=1}^{N} u_{ij}^{m}}$$
$$u_{ij} = \sum_{k=1}^{c} \left(\frac{\left\| x_j - v_i \right\|}{\left\| x_j - v_k \right\|} \right)^2 \quad 2\text{-}8$$

Different objective functions have shown a generalisation of the FCM algorithm. Many variations to the FCM have been suggested over the years because the traditional FCM could not operate excellently in the presence of outliers and inhomogeneity, which exist in real datasets. This results in centroids shifting toward outliers instead of the actual cluster centres (Qamar, 2014). These methods are suggested by using spatial data or assessing the diversity of the suspicion field or by altering the cost function to address limitations. These FCM variants are as follows (Gosain & Dahiya, 2016):

- Bias-corrected fuzzy *C*-means (BCFCM);
- Possibilistic fuzzy *C*-means (PFCM);
- Spatial fuzzy *C*-means (SFCM);
- FCM_S1 and FCM_S2;
- Fuzzy local information *C*-means (FLICM);
- Multi-dimensional fuzzy *C*-means (MDFCM);
- Weighted image patch-based FCM (WIPFCM);
- Kernel weighted fuzzy local information *C*-means (KWFLICM); and
- Strong FCM.

2.5.2 Gustafson-Kessel method

Gustafson and Kessel introduced the G-K method (GKA) in (Gustafson & Kessel, 2008) and (Babuška, 1998) to use an adaptive distance measure on cluster centres and data point covariance matrices to assess dissimilarity. Since the distance measure used in the GKA is in the form of the Mahalanobis standard, it may be taken into consideration that the GKA utilises ellipsoids to cluster data points. However, before iteratively calculating the cluster centres, the GKA assumes fixed ellipsoid volumes. The FCM and GKA are probabilistic fuzzy approaches to clustering.

2.5.3 Possibilistic clustering method

Krishnapuram and Keller (1996; 1993b) proposed the possibilistic clustering method (PCM) as a practical method for pattern recognition and data analysis. Pal et al. (Pal, Pal, Keller, & Bezdek, 2005) examined the PCM on Iris dataset and stated that they tried a variety of initialisations to get the PCM to find three clusters, but it constantly produced coincident clusters and eventually returned only two clusters regardless of how they initialised the PCM. Several PCM versions were subsequently suggested to improve the efficiency of the original method.

To enhance the PCM, Zhang and Leung (Zhang & Leung, 2004) introduced the fuzzy method to the PCM, thereby maximising the effectiveness of the possibilistic method. The typicality values calculated in the PCM would be low if the dataset contains many objects. Even though scaling appears to solve the problem of small values, scale values do not have additional data points' information. The fuzzy and possibilistic *C*-means method (FPCM) suggested by Pal et al. (1997) intended to combine the characteristics of the FCM and the PCM, and it was hence called the mixed *C*-means (MCM) method. The FPCM provides memberships and the likelihood for each cluster, along with the standard prototype point or centre of clusters.

Pal suggested the FPCM and PFCM, a combination of the PCM and FCM, to prevent coincident clusters. The drawback of the FCM about outlier sensitivity is avoided in the PFCM, and it prevents the coincidence cluster. Furthermore, the row sum restriction of the FPCM ($a + b = 1$) is eliminated, so that the PFCM is an excellent fuzzy rule-based system classification. However, there are four parameters to learn from the PFCM model, and finding the best four parameters in an uncertain environment is difficult. The PFCM has three outputs compared to other fuzzy and possibilistic approaches that mostly generate two outputs. The PFCM's outputs are U, a fuzzy partition or membership matrix;

T, a topicality matrix; and V, a set of c prototype points. Variables a and b in the PFCM describe the comparative significance of prototyping and membership in computation of centroids. While in the presence of noise, the PCM and PFCM work better, they are also not perfect. The PCM performs poorly in finding optimal clusters in noisy datasets, and the PFCM performs poorly if the dataset includes anomalies and if clusters are imbalanced in size.

Xie et al. suggested an improved PCM clustering method in (Xie, Wang, & Chung, 2008), which partitioned the original data into the primary cluster and the assisted cluster to prevent the coincident clustering.

The PCM can be used widely in the assessment and mining of big sensor data. Many datasets, however, suffer from incompleteness in big sensor data; that is, a dataset *X* may contain parameters that miss one or more ascribed values (Li, Gu, & Zhang, 2010). The PCM was unable to cluster such incomplete information sets in real time entirely. On the one hand, in incomplete datasets, the PCM could not measure the distance between two items, whereas uncompleted objects easily corrupt the accuracy of the PCM. On the other hand, in the presence of the substantial amount of data, meeting the actual-time requirement of clustering incomplete big data in the PCM is difficult. Zhang and Chen (Zhang & Chen, 2014) thus proposed a distributed weighted possibilistic *C*-means method (DWPCM) for clustering big incomplete data.

2.5.4 Robust fuzzy *C*-means

Tsai and Lin suggested that the traditional FCM should be changed to a fresh distance index called the FCM-σ by altering the distance measure used in the standard FCM (Kaur, Soni, & Gosain, 2011).

2.5.5 Type-2 fuzzy sets fuzzy *C*-means

All the above-mentioned fuzzy clustering methods have Type-I membership values. An FCM Type-II by (Rhee & Hwang, 2001) expanded the membership value of the FCM to the Type-II FCM as well. The initial concept in Type-II fuzzy sets fuzzy *C*-means (T2FCM) is that not all input should contribute equally to cluster centre computing. Instead, points with greater membership status need to dominate in processing cluster centres. The concept of constructing the fuzzy sets Type-II is merely based on the principle that the secondary membership function must produce the greater possible value, which should be higher than the lower possible value for the same Type-I membership value.

2.5.6 Kernel type-II fuzzy set fuzzy *C*-means

Kernel Type-II fuzzy set fuzzy *C*-means (KT2FCM) was proposed to solve the T2FCM problem by adding kernel, tangent, and Lagrangian methods, which still have the objective function for the T2FCM method as well as improved clustering in the presence of noise (Kaur et al., 2011).

2.5.7 Multi-kernel fuzzy clustering

Multi-kernel fuzzy clustering (MKFC) is an effort into the FCM approach that tackles the problem of the restriction of spherical shape clusters (Huang, Chuang, & Chen, 2012). It uses different kernels and instantly changes the kernel measures to protect the system from inadequate kernels and insignificant features.

2.5.8 Evolving fuzzy clustering method

The evolving fuzzy clustering method (ECM) was introduced in 2002 by Kasabov and Song (Kasabov & Qun Song, 2002), and it is named the first evolutionary online clustering (Ravi, Srinivas, & Kasabov, 2008). Evolving fuzzy clustering works in two

phases: online tracking and offline tracking. The number of groups in the one-way method is predicted interactively in the offline stage.

2.5.9 Intuitionistic fuzzy *C*-means

Meanwhile, in the case of digital images, it is not possible to comprehend precisely which pixel corresponds to which group. Some sort of hesitation exists regarding the concept of the affiliation function. This idea provokes Atanassov's (Krassimir Atanassov & Georgiev, 1993) concept of the higher fuzzy, labelled as an intuitive fuzzy set. The objective function of the intuitionistic fuzzy *C*-means (IFCM) is based on two central terms: the IFS objective function and the new fuzzy intuitionist entropy (IFE) (Chaira, 2010).

The IFCM expands the standard FCM by incorporating intuitive characteristics into the association and objective functions algorithm. It displayed better performance in comparison to available algorithms but was unable to cluster non-spherical separable samples effectively. Several IFCM versions were thus suggested to improve the efficiency of the original algorithm.

To manage the distance fluctuation in each group, robust intuitionistic fuzzy *C*-means (IFCM-sigma) was suggested by Kaur, Soni, and Gosain (Kaur et al., 2011), acknowledging a distinct distance measure to the IFCM. Therefore, the IFCM-sigma performs better than the IFCM in cases where the clusters are non-spherical.

Kernel intuitionistic fuzzy *C*-means (KIFCM) is another variation of the IFCM that implements the radial basis kernel feature to calculate the interval between the centre of cluster and items, was the recommendation of the KIFCM approaches (Gosain & Dahiya, 2016). The IFCM's precision was consequently enhanced to optimise the accuracies of intuitive FCM.

Another variation, radial basis kernel robust intuitionistic fuzzy *C*-means (KIFCM-sigma) incorporates the IFCM, the kernel function, and innovative measuring distances to determine the interval between the cluster core and objects. It promotes clustering or centre point processing without considering noise and anomalies, thereby enhancing the precision of the IFCM by solving problems of the IFCM and FCM-σ (Gosain & Dahiya, 2016).

2.5.10 Noise clustering

Dave's noise clustering (NC) method (Dave, 1991) indicates that a noise prototype distance from all points was defined as a constant value. The term "noise clustering" is implemented to designate the noise category to noisy data points. This method is designed for algorithms such as *K*-means or fuzzy *K*-means, which are objective-based function algorithms, to identify "excellent" clusters among noisy information.

2.5.11 Credibility fuzzy *C*-means

Credibility fuzzy *C*-means (CFCM) is another method that is explicitly suggested to operate with noisy records effectively. Instead of exactly identifying outliers, the NC and CFCM emphasise decreasing the impact of those outliers on the resulting clusters. The CFCM method proposed by Chintalapudi does this by applying a new matrix, named credibility, and it reduces the cluster computing effect that is caused by outliers. This demonstrates that the method strengthens centroid measurement. However, the centroid may not be accurate, since some outliers still exist in several clusters (Kaur, Soni, & Gosain, 2013).

2.5.12 Density-oriented fuzzy *C*-means

Density-oriented fuzzy *C*-means (DOFCM) (Kaur & Gosain, 2010) is another method developed to deal with outliers using a density factor, which is a group membership that would first evaluate the compactness of an object in a dataset by considering its

environment and then accomplish clusters considering actual data points. The DOFCM initially specifies outliers from a dataset with n proper clusters and one anomalous cluster consisting of noise and outliers. Then, it creates n clusters and one outlier cluster, which results in $n + 1$ clusters. In this process, the density method was used to define outliers, and cluster membership in the FCM was modified. The location of the centroids in the DOFCM model was consequently not affected by noise in the dataset.

2.5.13 Unsupervised fuzzy partitioning-optimum cluster number

The unsupervised fuzzy partitioning-optimum cluster number (UFP-ONC) method, introduced by Gath and Geva (Gath & Geva, 1989), interprets the membership values as likelihood estimates with ideal results. It performs well in conditions where the cluster shapes, density, and number of objects in each group are highly variable. However UFP-ONC is not suitable for application where memberships should be a demonstration of typicality or compatibility with a flexible restriction, as the memberships generated by this restriction are relative values (Zadeh, 1965, 1978).

2.5.14 Dynamic fuzzy clustering

An efficient method, called the dynamic fuzzy cluster, was introduced by Min Ji, Funding Xie, and Yu Ping (Ji et al., 2013) to actively group time series data by identifying main objectives and enhancing the FCM algorithm. It seems to operate by identifying the time series with uncertain category names and then dividing them into distinct clusters over the point of time. Compared to other existing algorithms, the primary advantage of this approach is that as time goes by, the characteristics of certain time series corresponding to specific groups can be largely found. The suggested algorithm may be implemented to fix some clustering issues in data analysis. A three-level dynamic fuzzy clustering (DFC) method includes: initial partitioning, a series of updating, and merging by using optimisation of a characterisation function, that is based solely on measures of

fuzziness in a set. Finally, in contrast to the conventional detection of separated preliminary clusters, the method can extract overlapping preliminary cluster boundaries when the feature space has ill-defined regions.

2.5.15 The conventional *k*-nearest neighbours based fuzzy clustering methods

A study on the development of fuzzy clustering methods from the conventional *k*-nearest neighbours (kNN) method was primarily driven by fuzzy clustering method's limitations. For instance, a clustering method centred on kNN is introduced by the authors in (C. S. Li, Wang, & Yang, 2010). This method is an ensemble approach that produces a data correlation matrix and then uses hierarchical clustering to accomplish final clusters.

To summarise the ensemble data, the algorithm produces the data correlation matrix and then pertains hierarchical clusters to accomplish the ultimate clusters. Another example is by (Weng, Jiang, Chen, & Hong, 2007), who proposed a distinctive cluster alignment method that creates a connection between fuzzy clusters by setting up a brand new cluster association pairing system and big fuzzy membership. Furthermore, a new method of clustering is suggested based upon the kNN model in (Guo, Wang, Bell, Bi, & Greer, 2003). It is similar to the kNN model, but it instantly determines the number of *k*. The model is developed by generating several training data representatives that are engaged in the clustering and whose size is much smaller than the total training content. In (L. Chen, Guo, & Wang, 2012), the kNN model is improved by creating a training method based on the cluster to discover the best representation set. In addition, the kNN based dynamic evolving fuzzy clustering (KEFCM) method conducted by (Abdulla & Al-Nassiri, 2015) addresses the issues of the price calculation, modifying fuzzy clustering, and the traditional complication of clustering with kNN. It implements the least square method to identify the centre of the cluster and its cluster border, along with the ED measure to assess the degree of membership that the KEFCM presents to the

neural fuzzy inference model as a pre-processor (Shubair, Ramadass, & Altyeb, 2014). The approach to the KEFCM concerns the dynamic evolution that differentiates it from the development of kNN discussed above. Moreover, the KEFCM takes place in the offline and online stages. The partitions are made by the KEFCM throughout the offline stage, while in the online step, it clusters upcoming data and sequentially upgrades the clusters to evolve.

2.5.16 Incremental fuzzy clustering

Incremental clustering was proposed to properly manage large data that cannot fit completely into the memory. Single-pass fuzzy *C*-means (SpFCM) and online fuzzy *C*-means (OFCM) are two representative incremental fuzzy clustering (IFC) methods that both extend the scalability of the FCM through piece-by-piece processing of the dataset.

The incremental clustering assumption is that datasets can be viewed and allocated to the clusters one by one. In this method, adding new objects would not significantly affect the current clusters. Moreover, to reduce the required storage, only cluster centres would be kept in the main memory.

The table below presents an overview of different fuzzy clustering.

Table 2-2: Overview of different fuzzy clustering methods

Fuzzy Set Type	Method	Reference	Year	Disadvantages	Advantages
Type I	FCM	(Bezdek et al., 1984)	1984	-Performs poorly in the presence of outliers -Sensitive to noise	-Easy to understand -Used in many disciplines
	PCM	(Raghu Krishnapuram & Keller, 1993a)	1993	-Sensitive to noise (A. K. Jain, Murty, Fynn, et al., 1999) -Highly sensitive to	-Helpful in outlier detection (Pal et al., 2005)

				initialisations (N. Pal et al., 2005) -Highly sensitive to parameter setting (N. Pal et al., 2005)	
	FPCM	(Pal et al., 1997)	1997	-When the number of *n* (number of data points) is large, the typicality values would be small (N. Pal et al., 2005)	-Does not suffer from the sensitivity problem that the PCM seems to exhibit
	PFCM	(Pal et al., 2005)	2005	-Performs poorly in the presence of outliers -Sensitive to unbalanced clusters (Kaur et al., 2013)	
	NC	(Dave, 1991)	1991	-Low accuracy in finding the right clusters -Reduces the effect of outliers but does not nullify them	Works well in the presence of noise
	FCM-σ	(Dave, 1993)	1993	Can only deal with linearly separable data points	Improvement over the FCM
	KFCM- σ	(Tsai & Lin, 2011)	2011	Can be effected by noise	Can handle non-linearly separable data points.
	CFCM	(Chintalapudi & Kam, 1998)	1998	Low accuracy in finding the right clusters	Works well in the presence of noise
	DOFCM	(Kaur & Gosain, 2010; Tasdemir & Merényi, 2011)	2010		Works well in the presence of noise

Type II	T2FCM	(Rhee & Hwang, 2001)	2001	Not suitable for non-spherical and complex clusters (Bhaskar N. Patel, 2012)	Generates better cluster centroids (Kaur et al., 2011)
	KT2FCM	(Tsai & Lin, 2011)	2011		-Generates better cluster centroids. -Performs well in the presence of noise (Kaur et al., 2011)
	GMKIT2-FCM	(Dinh Nguyen, Ngo, & Pham, 2013)	2013		Automatically determines the optimal number of clusters
	GT2-FCM	(Linda & Manic, 2012)	2012		Provides increased robustness in situations where noisy or insufficient training data are present (Linda & Manic, 2012)
	IT2-FCM	(Hwang & Rhee, 2007)	2007		Improved clustering result compared to the FCM
Intuitionistic	IFCM	(Kaur, Soni, Gosain, Soni, & Anjana Gosain, 2012; Z. Xu et al., 2008)	2008	-Fails to identify non-spherical clusters. -Only can deal linearly separable data points.	Low computation cost
	IFCM- σ	(Kaur et al., 2011)	2011		Ability to identify non-spherical clusters

	KIFCM	(Kaur et al., 2012)	2012		More accurate than the IFCM
	KIFCM-σ	(Kaur et al., 2012)	2012		Improvement over the KIFCM
	EKIFCM	(Lin, 2014)	2014		Combines advantages of intuitionistic fuzzy sets, kernel functions, and GA in actual clustering problems

For all fuzzy clustering methods that have already been reviewed, the number k of clusters must be pre-assumed. Since the number k is usually unknown, the method of finding the optimal k is important. This type of problem is usually referred to as cluster validity.

In some way, the utilisation of the above-mentioned evolving clustering methods is successful. The features of an efficient clustering method should be as follows: fuzzy clustering, dynamic change, small operational price, and low prior effort. However, this method still has not been fulfilled.

2.6 Clustering Time Series Data

Many real-world situations require that the data be classified into homogeneous groups, although prior knowledge about the structure is not available. This makes clustering a better-suited tool for pre-processing information in a complex information mining process and an unlabelled data assessment method.

In disciplines such as economics, environmental sciences, astronomy, physics, chemistry, biology, and many others, data that changes with time are a frequent occurrence. Since the data's trait values change as time passes, it is largely regarded as vibrant information, meaning that data is captured in continuous time intervals. As a result of enhanced information storage and handling capacity, real applications have been prepared to hold and retrieve information for a long period. Furthermore, because of infrastructure advancements that are prepared for all industries in the form of cloud computing and cloud storage, those industries have access to large data storage spaces at a significantly lower price. This enables businesses to keep all their information as time series over time. In addition, processing power is no longer an obstacle for analytics because of the development of new big data and cluster computing concepts. Time series data are then available in almost every industry, including finance, biomedical, physics particle analysis, biometrics data, air quality, and many more.

An unsupervised grouping of a collection of unlabelled time series into classes or areas, in which all components are placed in the same group, is called time series data clustering. Data from time series are important because of their growing popularity in various disciplines, such as commerce, scientific technology, banking, business, medical care, and administration (Liao, 2005). Each time series is a collection of data points that has been collected in time intervals, and it can be considered as a single item. Clustering these multi-dimensional items could be useful to determine whether an interesting pattern exists in the datasets. Numerous studies have been conducted to address time series clustering challenges, such as the following: identifying changes that occur in time series, recognising the character, and detecting anomalies and noises (Deshmukh & Hwang, 2019; Faloutsos, Ranganathan, & Manolopoulos, 1994; Kant & Mahajan, 2019; Yang, Lv, & Wang, 2006). Section 2.6.5 discusses further applications of clustering time series data to emphasise the significance of or need for clustering datasets in time series.

Time series datasets are quite massive, and they cannot be treated well enough by human examiners. Therefore, many users prefer to handle standardised data instead of bulky amounts of data. As a result, time series data is illustrated either by collecting data in clusters that are not overlapped or by classifying it as a hierarchical system of an abstract concept as a collection of similar time series clusters.

In time series studies, clustering and its branches, such as encoding, categorisation, and outlier recognition, are widely used as an exploratory method. Illustrating the cluster structures of time series as visible information (time series data visualisation) may enable a person to easily discover data frameworks, clusters, outliers, and other abnormalities in databases.

Definition: Time series Clustering: Given a dataset of *n* time series data $\mathrm{D} = \{\mathrm{F}_1, \mathrm{F}_2, .., \mathrm{F}_\mathrm{n}\}$, the process of unsupervised partitioning of *D* into $\mathrm{C} = \{\mathrm{C}_1, \mathrm{C}_2, .., \mathrm{C}_\mathrm{k}\}$, in such a way that homogenous time series are grouped together based on a certain similarity measure, is referred to as time series clustering. Then, C_i is called a cluster, where $D = \cup_{i=1}^{k} C_i$ and $C_i \cap C_j = \emptyset$ for i ≠ j (Aghabozorgi et al., 2015).

Clustering of time series is difficult, since its data is usually much larger than system memory and therefore archived on external hard drives. This slows down the clustering process significantly. The second issue is that data from time series is usually high-dimensional, making it difficult for many clustering algorithms to handle the data as well as slowing down the process of clustering. The third issue examines the metrics used to create the clusters by way of similarity (Keogh, Pazzani, Chakrabarti, & Mehrotra, 2000; Lin, Keogh, & Truppel, 2003). Similar time series are discovered based on their similarity values by use of a similarity metric. If the whole time series is utilised for the process of similarity calculation, it would be referred to as “whole sequence matching”. The method is nevertheless complex, as the time series database usually has noise and includes

anomalies and changes. These common issues posed a major problem for data specialists to find a similarity measure. Some of the most popular methods include the following:

- Hidden Markov models;
- Dynamic time warping;
- Recurrent neural networks;
- Dynamic Bayes nets;
- Constructive induction of temporal features;
- Extracting prototype examples; and
- Applying relational learning methods.

Clustering of time series can indeed be grouped into three categories: shape-level (whole time series) clustering whenever used on multiple different time series; structure-level (subsequent time series) clustering when administered on a single, long time series; and finally, time point clustering, which is the cluster of time points according to their proximity and the similarity of their equivalent values. The first two categories were mentioned in 2005 (Mörchen, Ultsch, & Hoos, 2005).

Furthermore, four major areas should be discussed in the study of time series clustering. These areas are illustrated in Figure 2-6 and discussed in the subsequent sections.

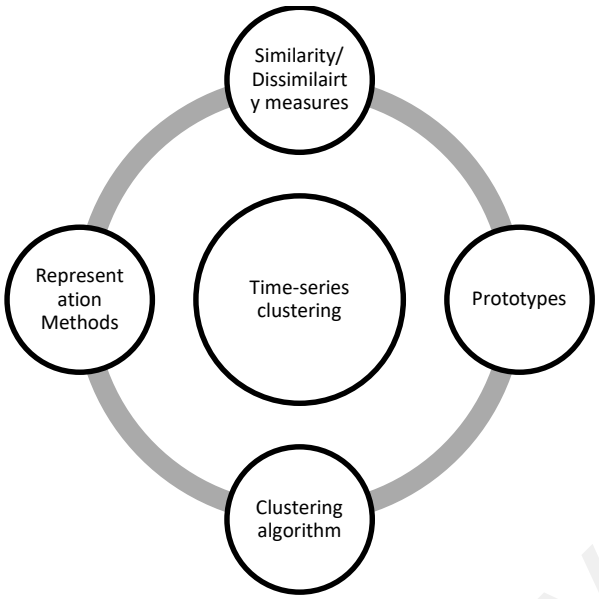

Figure 2-6: Four major areas of study in time series clustering

2.6.1 Major time series clustering approaches

Han and Kamber (2006) divided clustering for static data into five groups: density-based, model-based, partitioning, hierarchical, and grid-based approaches. Of these five main types of clustering methods, partitioning, model-based, and hierarchical methods have either been used directly or been modified for clustering time series data. Partitioning methods, on the one hand, have been extensively used because of their fast response, particularly in comparison to other methods. However, since the number of clusters must be appointed in advance, they are often more acceptable for clustering time series that are of a similar length because of their dependence on cluster representatives. On the other hand, in the hierarchical approach, it is not necessary for a user to define the number of clusters in advance, and this approach also has ideal visualisation in time series clustering. Moreover, hierarchical clustering has other advantages over the partitioning methods; for example, it can be utilised for time series data of unequal lengths, and it is superior when it comes to evaluate dimensionality reduction or distance metrics.

However, hierarchical clustering can only be utilised for small datasets because of its quadratic complexity. The use of model-based and density-based clustering is scarce for almost the same slow process problem and complexity reasons. In addition, model-based clustering approaches mainly rely on the user assumption. Therefore, the utilisation of methods based on density or model is uncommon. Several types of research have recently committed to strengthening methods by presenting new solutions based primarily on the wide variety of methods, such as hybrid clustering or multi-step clustering methods (Aghabozorgi et al., 2015).

Clustering of time series is among the interesting research on information retrieval concepts. Information collected in the time series is often massive, and there is the temporal ordering of elements of this type of data. As demonstrated in Figure 2-7, high dimensionality, temporal order, and noise are the main issues associated with the clustering of time series.

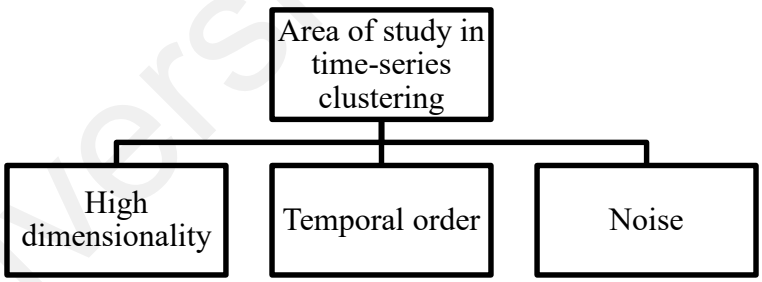

Figure 2-7: Area of study in time series clustering

Time series clustering is divided into three classes: temporal-proximity-based, if it operates directly on unprocessed data with respect to either degree or period of time; representation-based, when it acts indirectly by means of attributes driven from unprocessed information; and model-based, when it performs on the row information model (Rani & Sikka, 2012).

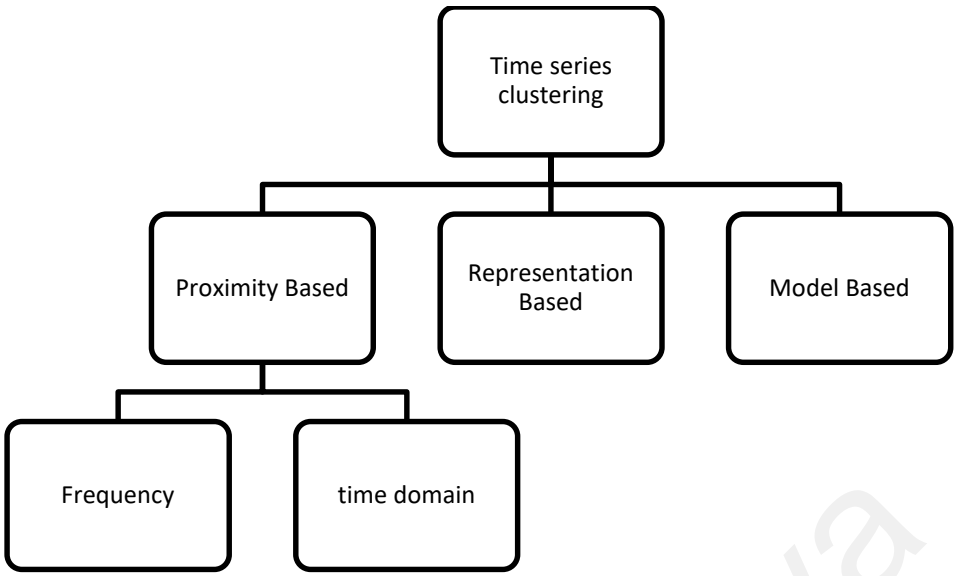

Figure 2-8: Grouping time series clustering based on how they treat raw data

Temporal-proximity-based method: This method generally operates directly with unprocessed time series records; therefore, it is known as the raw dataset method. The fundamental change is to replace the distance or similarity measure of static data with the proper time series measurement.

Representation-based method: Working directly with highly noisy raw data is not easy. This method first changes unprocessed information of time series to a parameter with low-dimensionality, and it then utilises clustering algorithms. One of the major problems in selecting representation techniques is indeed using a suitable and proper way of measuring similarity. Methods of representation and similarity measures are two main elements when handling time series information to accomplish efficiency and effectiveness. Time series information in principle has high-dimensionality, and it is costly to directly cope with information in its unprocessed format in terms of procedure and recording price. Therefore, developing classification strategies that can minimise dimensions while retaining the significant features of a specific information source is particularly necessary.

Model-based method: This method notices how every time series is generated on the basis of a pattern or possibility of occurrence. If the prototypes describing each series or the existing elements are similar after matching the model, then the time series will be regarded as similar. It relies on the individual's parameter expectations. Since time series samples are sometimes wide and may carry anomalies, the recommended methods must be thoroughly examined.

2.6.2 Time series clustering representation

Many methods for representing time series with reduced dimensionality have been proposed in the literature (Ding, Trajcevski, Scheuermann, Wang, & Keogh; Rani & Sikka, 2012). These methods include "discrete Fourier transformation", "single value decomposition", "discrete cosine transformation", "discrete wavelet transformation", "piecewise aggregate approximation (PAA)", "adaptive piecewise constant approximation", "Chebyshev Polynomials", "symbolic aggregate approximation", "indexable piecewise linear approximation", and "Symbolic Aggregate approXimation representation (SAX)" (Kadam & Appl, 2012).

The reduction of dimensions is a popular approach for most clustering methods in whole time series; it is advocated in the written manuscripts (Lin, Keogh, Lonardi, & Chiu, 2003; Lin, Keogh, Lonardi, & Patel, 2002; Niennattrakul, Srisai, & Ratanamahatana, 2012; Nunthanid, Niennattrakul, & Ratanamahatana, 2011; Shieh & Keogh, 2008) and is recognised as the representation of time series. Although a decrease in dimension translates into faster clustering, it is undeniable that more information would be missed. Finding a balance between precision and speed in representation methods is thus a contentious and crucial task. In other words, the rate of dimensional reduction is a highly subjective matter, and it depends on the problem to be addressed and the type of

time series. Given its power in representing, the focus in this dissertation is on the SAX between all these representation methods that may have their strengths and faults. It is implemented to decrease data dimensionality, and it is a symbolic representation of time series datasets that has been utilised by over 50 teams in several data mining studies (Hruschka, Campello, Freitas, & de Carvalho, 2009b; Jessica Lin, Keogh, Wei, & Lonardi, 2007).

Bagnall and Janacek (Bagnall & Janacek, 2005) indicate that in the presence of outliers, clustering precision is enhanced by the use of representative rather than unprocessed time series records. To summarise, SAX is as useful as any other familiar and commonly utilised classification method, such as discrete wavelet transform (DWT) and discrete Fourier transform (DFT), since less memory room is needed (Lin et al., 2007).

Symbolic aggregate approximation representation is indeed a two-step method that changes a time series to the PAA and relates the coefficients to symbols afterward. Consider $\bar{F} = \{\bar{f}_1, \ldots, \bar{f}_w\}$ as a discretised time series by PAA transformation. Then $\hat{F}$, where $\hat{F} = \{\hat{f}_1, \ldots, \hat{f}_w\}$, is defined by mapping the PAA coefficients to a SAX symbols, where a is the alphabet size (e.g., for the $alphabet = \{a, b, c, d, e, f\},\ a = 6$), and the alphabets in SAX are defined by "breakpoints". Based on the Keogh definition, a list of numbers $B = \{\beta_1, \ldots, \beta_{\alpha-1}\}$ is defined as "breakpoints" to determine the area of each symbol in the SAX transformation.

2.6.3 Distance measures approach

Clustering of time series relies heavily on the distance measurement. Different distance measurements are designed to define time series similarity. Agrawal et al. (1993) proposed the theoretical problem of time series similarity or dissimilarity, and it happened to be a primary theoretical concern for data mining research. Measures of similarity

applied in the study of time series could indeed be grouped into three main classifications: the distance of Lp-norm, statistical methods, and elastic measurements (Izakian et al., 2013). The choice of a similarity metric in the evaluation of time series data relies on the quality of the input data and the characteristics of the application. The Lp-norm distance may be used to compare two time series with the required predetermined features; L1 (Manhattan) and L2 (Euclidean) are the most commonly used examples of Lp-norm distances. Such distances may be used to make comparisons of time series in the original dataset or in dimensionality-reduced data (Izakian, Pedrycz, & Jamal, 2015).

The literature contains more than a dozen distance measurements (Rani & Sikka, 2012) for time series data similarity, including the "ED", "dynamic time warping (DTW)", "distance based on longest common subsequence (LCSS)", "edit distance with real penalty (ERP)", "edit distance on real sequence (EDR)", "DISSIM", "sequence weighted alignment model (Swale)", "spatial assembling distance (SpADe)", and "similarity search based on threshold queries (TQuEST)" (Kadam & Appl, 2012).

In (Izakian et al., 2013), (Izakian & Pedrycz, 2014a), and (Izakian & Pedrycz, 2014b), an improved model of the ED feature was discussed by authors for fuzzy clustering of datasets in time series. The initial time series data and different representation methods, as well as "DFT", "DWT", and "PAA", have been researched for clustering aims. To analyse the objects in the new representation, D'Urso and Maharaj (2009) changed the time series data by correlation coefficient, and they also changed the ED utilised to compare objects in the new representation. Thereafter, the changed data was grouped using an FCM approach. In time series information, a clustering-based method for noise identification was researched for comparing data in the new feature representation domain. The clustering-based method was explored by Izakian and Pedrycz for outlier identification in time series information (Izakian et al., 2015).

Choosing an accurate distance measure is a challenging task in time series clustering. This choice relies much on the characteristics of time series, their length, their representation method, and generally the intention for clustering. By far the most popular approaches for measuring similarity in the clustering of time sequence data are the ED and DTW. The focus would be on the common distance measures for continuous data.

The ED is the most popular measure for continuous data. Let F_i and F_j be two time series of length n. The ED between F_i and F_j is defined mathematically as follows:

$$dis_{ED}(F_i, F_j) = \sqrt{\sum_{i=1}^{n} (f_i - f_j)^2} \qquad 2\text{-}9$$

It is also possible to remove the square root phase of the above equation. The ED has been mentioned in many works as a simple, quick approach, and it has also been used to compare the efficiency of newly proposed distance measures. However, it may not always be the best option; the ED is highly dependent on the immediate problem and its time series features, and using it has some disadvantages in general:

1. The time series being compared should have precisely the same dimension or duration.
2. This metric is highly sensitive to small shifts in time (Keogh & Ratanamahatana, 2004; Ratanamahatana & Keogh, 2005). For instance, estimating sequence similarity in the following way is not accurate: $< adaa >, < aada >$.

Studies have revealed that the ED in time series is relatively reasonable in terms of classification precision (Lkhagva, Suzuki, & Kawagoe, 2006). This dissertation focuses on a comparable ED approach, which corresponds one similar item to another and is used in most research (Bao, 2007; Reinert, Schbath, & Waterman, 2000). In this research, the

FCM has been applied on time series data, and the ED measure has been used as a distance measure.

2.6.4 Time series clustering evaluation measures

The efficiency of a time series clustering method must be assessed using certain criteria. There are two distinct types of measurements based on known labels and unknown labels (Liao & Warren Liao, 2005; Rani & Sikka, 2012). This section discusses the method of evaluating the clustering. Keogh and Kasetty (2003) performed a noteworthy investigation on several time series research articles and concluded that the analysis of mining in time series must adhere to the following recommended principles:

- The verification of algorithms must be conducted in a different range of datasets (except when the algorithm is only generated for a particular dataset).
- The dataset that is used must be easily available or published, and a cautious design of studies must prevent bias in implementation.
- Wherever feasible, datasets and algorithms must be presented openly and publicly as free.
- A comparison of new similarity measurement methods must also be made with simple and reliable measurements, such as the ED.

Overall, assessing synthetic clusters with a lack of information labels is not a simple task and remains to be addressed. The interpretation of the cluster is subjective and mainly relies on the user who interprets the results and the application domain. For instance, the number and size of clusters, the concept of anomalies, and the clarification of the similarity between the time series are all principles that rely on the imminent condition and should be indicated personally. This has turned the clustering of time series into a significant issue in the field of information processing. Indeed, in the case where data

are tagged subjectively or are even synthesised by a generator, it is possible to evaluate the result using some measures.

The evaluation process in this research is based on the RI. The values of this clustering metric range from 0 to 1, where 1 indicates that the ground truth and cluster are equal. Therefore, greater values of criteria are favoured in this matter. Regarding clustering algorithms in time series, this section discusses the measurements of evaluation used in the different methods. In the following sections, the methods for evaluation of each suggested model, as illustrated in Figure 2-9, are discussed.

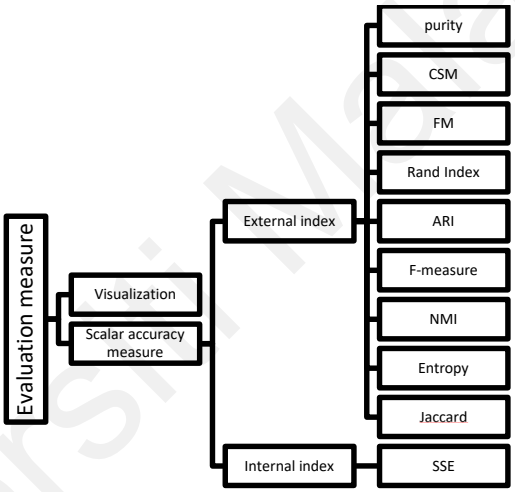

Figure 2-9: Hierarchy of measuring evaluation in the literature

In measuring scalar accuracy, a real number reflects the accuracy of different clustering methods. Furthermore, numerical methods used to evaluate different prospects of cluster validity are categorised into two groups: external and internal metrics.

2.6.4.1 External index

The external index is the most common approach of cluster evaluation used to evaluate the similarity of clusters that are created with externally obtained labels or ground truth (Halkidi, Batistakis, & Vazirgiannis, 2001). As there is ground truth, this measure is also regarded in the literary works as an “external criterion”, “external validation”, an

"extrinsic method", or a "supervised method". External validation indices are the concordance metrics between two groups or divisions. One will generally be a common group, which is also regarded as representative (e.g., model), and the other is achieved from the clustering procedure. Moreover, the ground truth is the desirable cluster almost always created by human specialists. Ground truth exists throughout this type of evaluation, and the matrix assesses the degree to which the clustering corresponds to it (Manning, Raghavan, & Schutze, 2008). Elaborate and extensive evaluations and reviews of some commonly used practises can be found in the literature (Aljarah, Mafarja, Heidari, Faris, & Mirjalili, 2019; Amigó, Gonzalo, Artiles, & Verdejo, 2009; Gan et al., 2007). Although several methods are available for clustering evaluation, there is no consensus and commonly acknowledged method for evaluating clustering approaches. Nevertheless, matching clusters with the ground truth is generally used as a method for external indices. Under these methods, some matrixes are addressed in various papers (Aghabozorgi et al., 2015):

- **Cluster purity.** Cluster purity is one way in which to measure the quality of a cluster solution (Zhao & Karypis, 2004). It is a clear and straightforward measure of valuation. In view of $G = \{G_1, G_2, \dots, G_M\}$ as a cluster of ground truth, $C = \{C_1, C_2, \dots, C_M\}$. Clusters are developed by a clustering algorithm based on evaluation; therefore, to measure the cluster purity of C with respect to G, each cluster is appointed to the most frequent class in the cluster group, and the validity of this application is subsequently assessed by dividing the number of properly appointed data by the total number of data in the cluster. A poor clustering has a value of 0 for purity, and an ideal clustering has a value of 1 for purity. Achieving high purity is not difficult when the number of clusters is high – especially when each item takes its very own cluster, purity will be 1. Accordingly, as a measure

of quality, one cannot rely solely on purity. Finally, purity was used in various studies to evaluate time series clustering (X. Wang, Smith, & Hyndman, 2006).

- **The Folkes and Mallow index (FM).** This metric is utilised to evaluate the precision of time series in clustering multimedia datasets (H. Zhang, Ho, Zhang, & Lin, 2006).
- **The Jaccard score.** Jaccard is a measurement utilised as an external index in different studies (Fowlkes & Mallows, 1983).
- **The RI.** This is a famous quality metric for the assessment of time series clusters that mostly measures the similarity between two partitions and demonstrates the degree of similarity between the clustering results and the model (Rand, 1971; Xiong, Wu, & Chen, 2009).
- **The ARI.** In a random cluster, the RI would not consider taking a 0 value (Hubert & Arabie, 1985). Accordingly, researchers have suggested an adjusted for the chance of the RI in which the performance is improved in respect to the RI and several other metrics (Milligan & Cooper, 1986; Steinley, 2004).
- **The cluster similarity measure (CSM).** The CSM is a simple cluster validity metric in the time series domain (Liao, 2005).
- **The F-measure.** The F-measure is a useful metric to assess the quality of any clustering method based on labels (Van Rijsbergen, 1979). It defines the degree to which each cluster is closed to a set of prototypes. This measurement has been applied in the clustering of time series data (Gullo, Ponti, Tagarelli, Tradigo, & Veltri, 2012) and for clustering evaluation in NLP (Kapitanov, Kapitanova, Troyanovskiy, Ilyushechkin, & Dorogova, 2019).
- **Normalised mutual information (NMI).** As indicated above, when there are many groups, high purity is not an advantage for the purity measurement. Normalised mutual information (Studholme, Hill, & Hawkes, 1999) has been used

as a measure of quality to choose between the clustering quality and the number of it (Cai, Chen, & Zhang, 2007). Furthermore, since this metric is normalised, NMI can be utilised for analyses of clustering approaches with a different number of clusters.

- **Entropy.** A cluster's entropy illustrates how spread-out the items are concerning the cluster (this has to be small). Entropy is a measure of item distribution in the generated cluster (Lin, Song, & Zhang, 2008; Rohlf, 1974).

In general, external evaluation is among the most common methods for assessing the quality of resulted clusters. However, in reality, the ground truth is not available for datasets for unsupervised tasks. In this situation, the internal index, which is discussed in the next section, will be used.

2.6.4.2 Internal index

When no ground truth is provided, the internal index is used to evaluate the quality of a clustering framework. It is recognised in literary works as “intrinsic”, an “uncontrolled method”, an "inner criterion", and "internal validation". Internal monitoring focuses on comparing data fitness between each cluster. If there is no ground truth, then it is usually better to use internal validity indicators, which assess clustering outcomes based on the characteristics of data and information in a data collection. It should be noted, however, that this metric can only make comparisons between various methods of clustering that were produced by identical measure, otherwise it allows cluster structure presumptions. Numerous inner variables exist, such as the “semi-partial *R*-squared (SPR) index, SSE, silhouette index, Hubert-Levin (*C*-index), Dunn index, separation index, Hartigan index, *R*-squared index, weighted inter-intra index, Krzanowski-Lai index, homogeneity index, Davies Bouldin, Calinski-Harabasz, and root-mean-square standard deviation (RMSSTD)” (Aghabozorgi et al., 2015). An important feature that characterises a

cluster's coherence is the SSE, and "better" clusters are expected to yield lower SSE values (Jiawei. Han, Kamber, & Pei, 2011). The SSE, which is the error of a distance-based metric, can be applied to assess the accuracy of clusters as the most common metric in various schemes (J. Lin et al., 2003; J. Lin, Vlachos, Keogh, & Gunopulos, 2004).

2.6.5 Applications of time series data clustering

Time series clustering has been used mainly to discover important features in data. The aim of time series clustering is twofold. First, it has been used to find the most frequent time series features and cluster them into groups that define overall patterns in datasets to anticipate the formation of a specific time series based on the corresponding group (J. Lin et al., 2003), such as gene expression profiling (Wu et al., 2019) and segmentation (Kim et al., 2019). Second, time series clustering has been used as a method for finding patterns that have surprisingly occurred in datasets, for example outlier identification (Mishra & Chawla, 2019) and trends discovery (Ohana-Levi et al., 2019).

In short, in real-world problems, clustering time series can be of interest for the following issues:

- Recognition of dynamic time series changes;
- Clustering-based prediction and recommendation;
- Discovery of patterns; and
- Detection of the anomalies.

Various applications of time series have been studied in geography, as large time series have been and still are to be obtained using contemporary data acquiring methods (Ji et al., 2013). Furthermore, an instant demand exists for new operational and productive methods of mining unfamiliar and unpredictable data from relatively large geographic datasets with wide dimensions and variations. Spatial information analysis and geometrical development have appeared as ongoing areas of study to tackle these

problems (Jeremy & Diansheng, 2009). Clustering spatial data in geography is a fundamental and vital issue. Some reported cases of time series cluster assessment focused on environmental information. For example, in Bode et al. (Bode, Schreiber, Baranski, & Müller, 2019) and in Meger et al. (Méger et al., 2019), time series clustering is applied to labelling building energy data and used for capturing crustal deformation respectively. Other instances can be discovered in medication, scientific research, computer science, and many more. New approaches were introduced by Wang and Chen (N. Y. Wang & Chen, 2009) and Killick et. al (2010) in which they presented a mathematical method centred on the means clustering method for managing ecological collected records to assess and estimate the power provided at a specified location by various green energy sources. Moreover, the clustering of developing countries was researched by Alonso et al. (Alonso, Berrendero, Hernández, & Justel, 2006) based on contemporary CO2 pollution data. Other instances can be found in, inter alia, medication, science, and financing.

2.7 Fuzzy Clustering Methods for Time Series Data

Conventional time series analyses have several shortcomings, namely, the presence of noise, the need for expert judgement for enhancing the models, and an overall challenging modelling operation. Fuzzy time series clustering attempts to bring simplicity to modelling and enhances conventional time series approaches by reducing the influence of noise and managing the instability (Duru & Bulut, 2014).

The fuzzy time series method can be illustrated through various aspects, such as the "rule-based forecasting method", the "rule of thumb solution", an "educated guess method", "pattern recognition", "time series clustering", or "heuristic modelling." All of these concepts denote the fuzzy version of time series and display their repackaging of information and uncertain case-based traits (Duru & Bulut, 2014). Furthermore,

approximate reasoning for automatic control systems is by far the most important achievement of fuzzy logic.

The most important feature provided by fuzzy logic is approximate reasoning for automatic control systems, according to Zadeh (1979). The fuzzy time series is an excellent approach of approximate reasoning by using similarities, information, and the limitation of data within a specified framework.

In clustering fuzzy time series, there are two crucial elements: selecting the correct distance and improving the degrees of membership. Coppi et al. (Coppi, D'Urso, & Giordani, 2006) and D'Urso (2005) briefly presented the two primary reasons for using the fuzzy strategy in time series clustering:

1. Awareness is increased to capture the information that characterises the pattern of the time series. Conventional clustering methods are likely to neglect this fundamental structure in many areas because of cycling or changing dynamics. However, fuzzy clustering can handle changes from one time state to another in patterns or positions that are not clear and not specific to a certain time.
2. If the time patterns are not too different, then there would be higher flexibility to define the centroid of the time series. In this case, the fuzzy cluster description enables the identification of the fundamental structures if they are expected to be within the specified time sequence.

The use of fuzzy methods in time series is more reasonable than non-fuzzy methods because of the dynamic characteristics of time series in many real-world applications, leading to changing patterns over time (Maharaj & D'Urso, 2011).

Dun (1973) proposed the FCM, and it is continuously upgraded by many researchers. It is an improvement on *K*-means whereby each record can become a part of

different groups with a membership degree. To address the issue of time series that are not aligned and to customise the FCM method, the cross-correlation clustering (CCC) method has been suggested by Höppner and Klawonn (Höppner & Klawonn, 2009). It may be applied for short-period time series (full system clustering) as well as to cluster succeeding time series (STS). Levet and others (Möller-Levet, Klawonn, Cho, & Wolkenhauer, n.d.) introduced a new FCM method, especially suitable for short time series and those with unevenly spaced sampling points. Another approach of fuzzy clustering relies on an autocorrelation characteristic of the time series. In this approach, a time series does not belong to a single cluster but to separate groups with different membership values (D'Urso & Maharaj, 2009; Ji et al., 2013). Furthermore, to cope with more complex datasets, Kannan et al. (Kannan, Ramathilagam, & Chung, 2012) suggested another extension of FCM clustering methods, named the quadratic entropy-based FCM. This current research presents a new approach for using a fuzzy membership matrix to develop a clustering method that enables heuristic post-pruning of data after clustering.

2.8 Critical Discussion

Based on the literature regarding fuzzy clustering, new methods are developed to address shortfalls to ultimately improve the quality of fuzzy clustering. One of the issues after the main introduction of the FCM by Bezdek (Bezdek et al., 1984) was the sensitivity to outliers and noise. Some researchers opted to modify the algorithm to tackle the noise issue. For instance, Dave (Dave, 1991) introduced the noise clustering method to reduce the effect of noise. The method suggests defining a noise prototype, which is a universal entity, in a way that it is always at the same distance d from every point in the dataset. This means that all the points would have an equal a priori probability of belonging to a noise cluster. The idea is that as the algorithm iterates, it will discover the noisy data, add them to the noise cluster, and keep the main partitions clean of noisy data. However, the

shortfall of the method is the specification of the noise cluster distance d. The pre-specification of d is not easy in practise, because in most cases, information to decide the value of d is not available. Moreover, the value of d would be different for different problems, and it would be based on the statistical parameters of the dataset.

Other methods with a focus on algorithmic and conceptual changes are the PCM and PFCM, which performed better in the presence of noise when compared with the FCM; nevertheless, they have their own weaknesses. On the one hand, although the PCM deals better with outliers, it fails to find optimal clusters in the presence of noise. On the other hand, once outliers are present, the PFCM fails to produce proper results if there are two clusters that are highly unalike in size.

Apart from the above-mentioned methods, other researchers decided to add additional fuzzy concepts to deal with the noise and outlier problem. For example, T2FCM is based on Type-II fuzzy sets, which enables some data to contribute more in computing appropriate cluster centroids; however, it fails when the data structure is non-spherical and complex. The intuitionistic fuzzy approach by the IFCM method is another concept of fuzzy sets that has been introduced to improve clustering. The IFCM improves the FCM by adding an intuitionistic feature to membership and objective functions, and it mainly improves cluster computation compared to existing algorithms; however, it could not resolve the clustering issue with non-spherically separable data. These methods have been receiving more attention from the big data community in recent years and across various disciplines. For example, Shukla and Muhuri (2019) are utilising Type-II fuzzy uncertainty modelling in gene expression datasets, and Kousar et al. (2020) are utilising Type-II fuzzy logic to improve hierarchical clustering to utilise in mobile wireless sensor networks. The intuitionistic fuzzy approach has received similar attention, and various new methods have recently been introduced, such as interval intuitionistic fuzzy

clustering (Lin, Duan, & Tian, 2020) and the interval intuitionistic fuzzy clustering algorithm (Lin et al., 2020). These methods are especially popular in the medicine discipline for MRI brain image segmentation (Kumar, Agrawal, & Singh Kirar, 2019; Kumar, Verma, Mehra, & Agrawal, 2019).

Another issue with distance-based clustering algorithms is that they utilise the ED and are consequently bound to the limitations of this measure, and they can tackle mainly spherically separable data. Therefore, another group of works focused on improving existing algorithms to be able to work with more types of data. Methods referred to as robust and kernel-based approaches are primarily the focus in this type of issue. Methods such as the FCM-δ, K2FCM, KIFCM, and KIFCM-δ are from this type of study.

As per our discussion above, works in the fuzzy clustering area can generally be divided into three main categories:

1. those focused on new concepts, approaches, and algorithms for better results in fuzzy clustering;
2. those focused on similarity or distance measures to make existing algorithms more robust to deal with additional types of datasets; and
3. those focused on utilising more fuzzy concepts in clustering algorithms.

The current research work is opening up another frontier in research into fuzzy clustering that has not been carried out before, and that is the following:

4. utilising the fuzzy membership matrix to improve the performance of the fuzzy clustering method.

2.9 Neutron and Gamma-Ray Discrimination

The main issue in the detection of neutrons is the discrimination of neutrons from gamma rays. Identifying emitted protons that are produced by fast neutrons is the most popular approach of neutron detection, and the variety of neutron detector applications is growing rapidly. Neutron detectors are currently used in nuclear research, nuclear medicine implementations, neutron imaging methods, and safety. They are also applied in different science domains, such as nuclear physics, aviation, medicine, and security (Yanagida, Watanabe, Okada, & Kawaguchi, 2019). Moreover, neutrons precede any gamma-ray diffusion. Together with gamma rays, neutrons radiate from nuclei. The time-of-flight (TOF) method is one of the well-recognised methods for discrimination between neutrons and gamma rays (Akkoyun, 2013), and the detector signal's PSD is another method used in various disciplines of nuclear physics.

Figure 2-10 is a block diagram of the detector system and indicates two channels that are used for traditional TOF calculations.

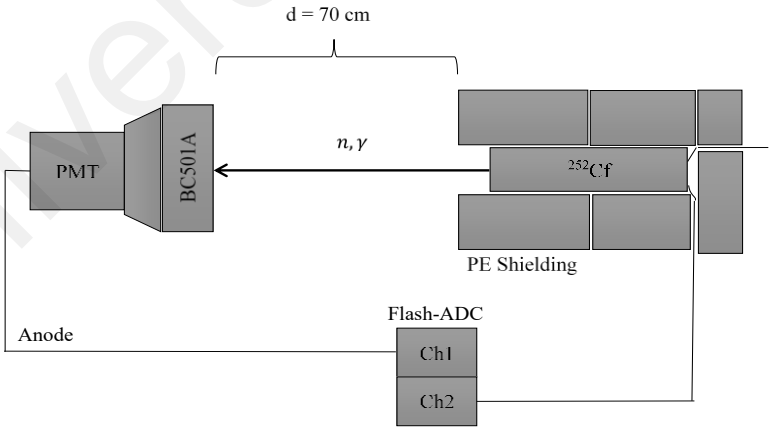

Figure 2-10: Block diagram of a scintillator and devices

To calculating the TOF, two separate datasets are required: one coming from the scintillator (Ch1) and the other from the neutron source (Ch2). The TOF is defined as the time that it takes the particle to fly distance d from the neutron source to the scintillator

detector. Figure 2-11 depicts the calculated TOF for our dataset. In the figure, the peak denotes gamma rays, while the bumpy part indicates neutrons.

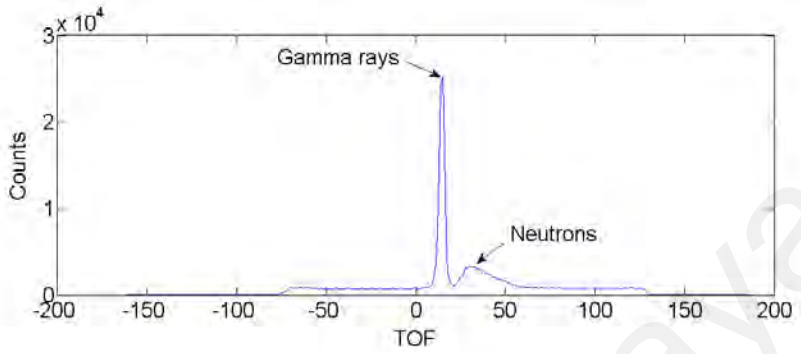

Figure 2-11: Traditional TOF discrimination method

These recoil protons are extensively detected by organic scintillators. Fast neutron photons in organic scintillators generate recoil protons via (*n*, *p*) elastic dissipation, and a recoil proton's maximum energy level is equivalent to the neutron's energy (Budakovsky et al., 2007). Furthermore, a wide variety of radiation has frequently been detected and diffracted by organic scintillation detectors. Neutron detection has numerous applications in the fields of atomic material control and national security, amongst others. Neutron detection utilising organic scintillators is lately being used for tomographic imaging. Given that organic scintillators are indeed responsive to gamma-ray particles, there is a need for a method to discriminate the pulse shape between neutrons and gamma rays (Yousefi et al., 2009).

Some organic scintillators are prime examples of this notion, which is implemented via PSD, such as stilbene and many industrial liquid scintillators. This is because of the significant variations caused by different radiations in the slow element trigger in these devices. While using organic scintillators as neutron sensors, PSD is crucial because all neutron areas are accompanied by a gamma ray. With different levels of success, several

PSD methods have been evaluated; the charge comparison method (Brooks, 1959) and the zero crossing method (Alexander & Goulding, 1961) are the most common. In analogue instrumentation, both methods are primarily applied, often in devoted electronics systems or nuclear instrument units (NIMs). As a result of the advancements in the digital domain, electronic devices have become available, and in recent years, both of the aforementioned methods have consequently been implemented for neutron and gamma-ray discrimination and have become industry norms to evaluate the new approaches, such as the correlation technique (N. V. Kornilov et al., 2003) and the method of curve fitting (Guofu Liu, Joyce, Ma, & Aspinall, 2010; Marrone et al., 2002).

Improving fast analogue in digital adapters and utilising digital devices allow analogue methods to be applied in the field programmable gate array (FPGA) as well as in the implementation of recent PSD methods for digital image processing. The optimal PSD filter described by Gatti (E Gatti, 1962) nearly 40 years ago was digitally implemented by Barton et al. (Barton & Edgington, 2000). Later, in (Back et al., 2008), the optimum filter method of the Gatti approach was implemented in a liquid scintillator to differentiate between α particles and β particles. Gatti's approach focused on the calculation of a G variable, which was positive for α electrons and negative for β pulses. This variable demonstrates the probability of the coming waveform being generated by alpha or beta ionisation (Yousefi et al., 2009).

Inorganic crystalline materials are the main tool in medical applications, and they have also been highly common in detectors in the security domain since the middle of the last century. Compared to most inorganic scintillators, organic scintillators are distinguished by a high light yield and fast decay times with an appropriate emission wavelength compared to the most conventional photodetectors, such as the photomultiplier tube (PMT). Anthracene is the most conventional organic scintillator and is available in a

standard ionising radiation measurement textbook (Knoll, 2010); this tool is considered to be a standard scintillator. Thanks to recent technological developments, several solid-state scintillators have been proposed, for instance the introduction of a new chemical composition such as fluorocarbon materials (Hamel et al., 2014) and new structural challenges such as microsphere materials (Santiago, Bagán, Tarancón, & Garcia, 2014; Yanagida, Watanabe, & Fujimoto, 2015). On the other hand, liquid neutron scintillators offer the opportunity for the PSD of proton particle emissions. This is of significance because neutron identification often takes place in the presence of a powerful photon context.

Dissociation is based on the reality that relative to photon-induced occurrences, incidents induced by neutrons generate pulses that have a higher portion of the longer scintillation component. A range of distinct methods was explored to translate this significant difference in sensor signal shape into a frequency for discrimination. The most commonly beneficial methods are the integration or delayed charge approach (Adams & White, 1978; Alexander & Goulding, 1961) and the zero crossing method (McBeth, Lutkin, & Winyard, 1971; Sperr, Spieler, Maier, & Evers, 1974), both of which have been applied for digital and analogue devices (Kaschuck & Esposito, 2005; Kornilov, Fabry, Oberstedt, & Hambsch, 2009; Moszyński et al., 1994; Söderström, Nyberg, & Wolters, 2008; Wolski et al., 1995). For this form of scintillator, although the achieved discrimination performance of these methods should be near to the peak, the disadvantage is that the input factors, such as areas of integration or shaping periods usually needs cautious manual tuning. This is particularly unsatisfactory in the case of large scanner devices with several channels. These parameters are dependent on each specific sensor and may change based on the experimental circumstances. The main pulse shapes that are driven experimentally can be applied to modify these variables, as mentioned above,

because the ideal adjustment depends on the primary pulse shapes (Gatti, Martini, 1962; Roush, Wilson, & Hornyak, 1964; Söderström et al., 2008).

The resulting composite curve for the large number of liquid organic detectors consists of two incremental disperses of the scintillator's fast and slow particles. Since the ratio of beam that appears in the slow element mostly depends on the type of current component, this reliance may be used to distinguish between different types of radioactivity. A liquid photon detector is among the most common radiation identification tools as it can be formed for a particular action according to the necessary magnitude. Another benefit of using this detector would be its useful PSD features and quick processing time (Yousefi et al., 2009). To create a molecular framework where unbinding *p*-electrons are expected to stimulate immediate radiation, organic scintillators of liquid and plastic form are generated. Such activation can contribute to stimulating *p*-electrons from the base state (S_0) to one of the stimulated regions $(S_1, S_2, S_3, etc.)$. Later disintegration from this position results in the release of a ray with visible light and that occurs just a few milliseconds after excitation. Furthermore, the fluorescence intensity of an organic scintillator decays exponentially (Birks & Firk, 1965). Different decline mechanisms exist when an irritated *p*-electron experiences a spin shift from the spin 0 singlet situation to the spin one triplet condition, leading to a decay of T1-S0 and a larger-range wavelength in comparison to the light emitted from fluorescence (phosphorescence) (Birks & Firk, 1965).

In a T_1 state, a π-electron may achieve adequate energy to go back to the S_1 position. This energy can be thermic, or the two π-electrons on the T_1 position can possibly interact and leave one of them in the position S_0 and the other in the state of S_1 along with the particle emission (Brooks, 1979). The next decomposition of the S_1 atom transmits late

ultraviolet light with the same characteristics of fluorescence, with the exception that strength does not decrease significantly (Miller, 2017).

Since the triplet intensity is calculated by the level of power failure of the incident particle, heavier particles display a significantly higher rate of power loss and generate late fluorescence, resulting in subatomic particles that decay much slower than those of lighter rays. The difference between the waveforms arising from the collision of heavy elements in the atomic detector and those arising from the contact of light electrons and photons has been used in PSD and makes it possible to determine the radiological category (Yousefi et al., 2009).

With effective applications across many areas, artificial neural network (ANN) platforms and the fuzzy systems (FS) method have appeared as advanced methods in recent years. They are especially efficient as pattern identification instruments and can therefore be used to classify neutron and gamma-ray incidents from the results conducted by organic liquid scintillation sensors using ANN. Furthermore, neural networks have recently been utilised to detect neutrons (Söderström et al., 2019). D'Mellow et al. (2007) viewed a computationally simple pulse gradient analysis (PGA) approach to discrimination (Guofu Liu et al., 2010). This method offers instant, digital representation of contexts in which both neutrons and gamma beams simultaneously exist. The effectiveness of the PGA method was analysed against the digital application of the classical charge comparative method, which demonstrated that the PGA method improves discrimination (Bao & Yang, 2008). Moreover, Ronchi et al. (2009) used an ANN for discrimination, and they experienced large performance advantages in comparison to other discrimination methods such as the Q_1 - Q_2 method and the Q-IRT method. In addition, Savran et al. (2010) deployed a combination of fuzzy logic and clustering, which is a machine learning method, to introduce a new approach for neutron and gamma-ray

discrimination. The robustness of this method stems from the fact that it is an unsupervised method and there is no need to assume the particular shapes (Joyce, Aspinall, Cave, Georgopoulos, & Jarrah, 2010; Guofu Liu et al., 2010).

All these PSD methods use the signal's time-domain characteristics; for example, the application of the charge comparison method usually depends on the inclusion of the wave over two distinct periods, and the PGA approach is centred on comparing current sample peaks in the pulse running border . Yousefi et al. (2009) suggested a special PSD method capable of spotting neutrons and electrons in liquid detectors based on converting pulses. In this method, because the features of both low-frequency particles are obtained at 512 and 1,024 modules, the discriminatory matrix is less responsive to high-frequency pulses than the PMT-induced pulse signal frequency, which is visible in the frequency range of pulses. Furthermore, observational findings indicate that the wavelet-based method enhances the reduction of the overlap of neutron and ray occurrences as compared to the PGA method, reflected by figure of merit (FOM) increases. However, the wavelet-based PSD method's operational computation is harder than the PGA matrix and is therefore not as appropriate for immediate discrimination. In addition, the method's efficiency may decrease at processing speeds aligned with the current system, in contrast to the comparatively endless headroom for the processing of the digital monitor spectrum analysers from which the spectrum was created (Guofu Liu et al., 2010).

Conventional analogue pulse discrimination methods are less flexible and more time consuming than novel digital approaches. The main problem with traditional discrimination methods is that they require various sources of data to perform the discrimination task. By clustering, pulses originating from the scintillator detector will be separated based on their shapes. No dataset from the neutron source (Ch2) is consequently needed, so it will provide the flexibility to instantly distinguish neutrons' rays from

gamma particles. Moreover, there is no need to wait to collect all datasets and start clustering, each pulse can be clustered as soon as it reaches Ch1 output. Additionally, digital technology affords some significant privileges, such as clarity of energy, increasing throughput, smaller size, easy upgrading and updating, automatic critical adjustments, multitasking operations, and automatic testing and verification. Table 2-3 summarises the discrimination methods for neutron and gamma rays. As demonstrated in the table, early methods were analogue; the emergence of digital technology opened a new door to digital methods, and since then, many methods have been introduced to deal with digital signals.

Table 2-3: Discrimination methods for neutron and gamma rays

Method	**Reference**	**Digital/Analogue**
Charge comparison method	(Brooks, 1959)	Analogue and digital
Zero crossing method	(Alexander & Goulding, 1961)	Analog and digital
Frequency gradient analysis (FGA), based on Fourier transform. Pulse-shape discrimination (PSD) method.	(Guofu Liu et al., 2010)	Digital
Correlation method	(Kornilov et al., 2003)	Digital
Curve-fitting method	(Marrone et al., 2002)	Digital
Artificial neural network (ANN)	(Liu, Aspinall, Ma, & Joyce, 2009)	Digital
Pulse gradient analysis (PGA)	(D'Mellow et al., 2007)	Digital

In the present work, the digital PSD using fuzzy clustering was investigated. The aim was to develop general-purpose PSD methods for particle identification, loosely coupled with detector or particle-type characteristics. Since the method is not targeted specifically to liquid scintillators, it is commonly used and can also be implemented in other types of sensors. The dataset for neutron and gamma rays used in this study is a real dataset provided by Savran et al. in (Savran et al., 2010).

2.9.1 Clustering for finding principal pulse shapes

In nuclear physics, a critical procedure is the differentiation of neutron and gamma-ray pulses. Preceding methods of analogue or digital discrimination are predominantly based on the analysis of physical or chemical variables. and they involve distinct information sources, while there are signals from a scintillator in the clustering information source, which is coupled with a digitiser. The major problem with traditional methods of discrimination is that different sources of data are needed to perform the task of discrimination. Based on their shapes, pulses from the scintillator detector will be differentiated by clustering. Therefore, no dataset from the neutron source (Ch2) is needed, so it will offer the flexibility to differentiate neutron rays from gamma particles instantly because there is no need to wait to obtain all datasets and begin clustering, and almost every pulse can thus be clustered as soon as it reaches Ch1 output. In many scientific areas, clustering is a well-known method for the discovery of knowledge, and researchers in nuclear science have recently begun to take advantage of machine learning and artificial intelligence strategies for particle discrimination (Akkoyun, 2013; Doucet et al., 2018; G. Liu et al., 2009; Ronchi et al., 2009; Söderström et al., 2019).

Particle classification by the evaluation of the pulse shape from a pulse detector is an efficient instrument in many fields of nuclear detection. The performance of such a detection (if any) is fundamentally limited by the sensor characteristics being used and

the various pulse shapes of varying emitted rays. Each type of particle relates to a certain prototype signal, and individual signals produced by a particular sort of particle match this model pulse shape, except the noise and statistical changes. These prototype signals related to various types of particles are known as the principal signals or prototype signal shapes. The role of the applied signal evaluation is to produce an image that somehow reflects the shape of a specified pulse so that distinct incident signals can be differentiated in each case. The way in which to identify (or measure) this amplitude might not be specific, and various methods may lead to different distinguishing qualities. In several instances, an appropriate signal shape amplitude is generated by evaluating integrals in different places of the signal coming from detectors, and variables (e.g., integration areas) need to be customised personally. Awareness of the principal signal shape for a specified incident signal and a specified sensor enables optimisation of the PSD's algorithm or may assist in modifying its parameters. The purpose of this research is to demonstrate that the proposed EFCA is a useful way in which to discriminate neutron and gamma rays.

This research demonstrates how to use the proposed EFCA to determine such prototype pulse shapes based on several digitised pulses. The number of various prototypes is apparently the only entry variable; therefore, the method provides an unsupervised approach to evaluating the primary pulse patterns.

CHAPTER 3:

METHODOLOGY

3.1 Introduction

This chapter elaborates on the methodology employed in this research. The motive and summary for time series clustering of the proposed evolving fuzzy method are discussed. The chapter also describes how the study is designed to accomplish the research objectives stated in Chapter 1. Furthermore, a description is provided regarding how the proposed method was evaluated as a general-purpose method for time series clustering, and its performance in this study's focused area, which is neutron and gamma-ray discrimination, is discussed. Finally, a chapter summary concludes this chapter.

3.2 Research Strategy

The framework of the research methodology for this thesis is depicted in Figure 3 -1.

Figure 3-1: Research strategy

3.2.1 Reviewing related works

An analysis of current approaches provides a broader view of clustering issues; therefore, the features and attributes of different clustering approaches were analysed based on a systematic review and investigation of different time series clustering methods in a discrimination task to represent their output and compare their quality of clustering.

Reviewing the literature in this research was twofold. First, since neutron and gamma-ray discrimination can be viewed as a clustering problem, the literature on current approaches to fuzzy clustering and time series fuzzy clustering required evaluation to

gain a broad view of the issues with fuzzy methods. Second, various methods used to discriminate neutrons and gamma-rays were studied, along with the data mining methods utilised directly in the discrimination of neutron and gamma-rays, and the possibility of using the proposed evolving fuzzy clustering approach (the EFCA) were investigated in this matter.

The effects of various similarity factors on the outcomes of distance-based clustering methods were examined, compared, and evaluated because they are crucial in distance-based partitioning clustering methods, regardless of whether they are soft approaches, such as the FCM (Bezdek et al., 1984), or hard clustering approaches, such as *K*-means or *K*-medoids.

3.2.2 Problem formulation

The problems with clustering methods are clearly explained in the literature review. Clustering is an unsupervised machine learning method that is used both individually and as a part of the pre-processing stage for supervised machine learning methods. Given its unsupervised nature, clustering results have less accuracy compared to supervised learning. However, most of the time-clustering results are not accurate enough and can result in deficient models. Therefore, it is desirable to have a smaller dataset (e.g., 80% or 50% of the primary dataset) but with a reasonable clustering accuracy, for instance 80%, instead of having all datasets clustered with a low accuracy of around 55%. By considering neutron and gamma-ray pulses as time series data, PSD shares the same set of challenges, solving these problems motivated the researcher to conduct this study to ultimately improve neutron and gamma-ray discrimination. This thesis thus aims to increase the accuracy of clustering approaches for the neutron and gamma-ray discrimination problem.

3.2.3 Defining the research objective

Based on the above-mentioned problem, the research objectives are as follows:

1. to develop a new method for clustering that is more accurate for neutron and gamma-ray pulses;
2. to evaluate the capability of the suggested method for improving the accuracy of the clustering; and
3. to improve the performance of neutron and gamma-ray clustering (discrimination).

To assess the capability of the proposed method to improve the accuracy of the clustering, the purpose of this study is to suggest an evolving fuzzy clustering method that can tackle the problems of neutron and gamma-ray discrimination using a fuzzy clustering method. A clustering model has been introduced in this study to achieve the objectives, as mentioned earlier. The proposed model is described in the following section.

3.2.4 Proposed model

This study seeks to introduce a new perspective in clustering by defining an approach for data pruning, namely, the EFCA, which enables clustering to use multiple sets of prototypes instead of only one set to improve clustering accuracy. This approach has the potential to be used independently or as part of the pre-processing stage to prepare purified data for the training step of a supervised learning approach. The EFCA utilises the fuzzy membership concept to break down clustering into epochs instead of running the clustering on all data at once. In some cases, for supervised learning, having a smaller subset of high-accuracy tagged data is preferable to having all dataset tagged with low accuracy. The EFCA's "epoch cut" enables post-pruning ability to eliminate obscure data points, which results in higher clustering accuracy.

For this purpose, the proposed approach must utilise an appropriate distance measure. This study therefore first analyses and evaluates the distance measures available in the literature for continuous data to discover the best-performing one.

3.2.4.1 Stage 1: Selection of similarity or dissimilarity measure for continuous data

Similarity measures are not limited to clustering; however, many data mining algorithms effectively use similarity measures to some extent. It must be noted that they have an essential impact on clustering performance and that they are worth researching. In this section, experimental research was conducted in different fields to demonstrate the impact of each distance measurement on data analysis, and the outcomes obtained by various distance matrices were compared and evaluated. While it is generally not practical to introduce a "best" measure of similarity or the best performance measure, a comparative study might illustrate the performance and behaviour of the measures. A variety of continuous data similarity measures were examined on low- and high-dimensional continuous datasets to define and evaluate the accuracy of each similarity metric in separate datasets with different dimensionalities, using 15 datasets (Dheeru, Dua and Karra Taniskidou, 2017; Fu & Medico, 2007; Gionis, Mannila, & Tsaparas, 2005; Veenman, Reinders, & Backer, 2002; Zahn, 1971), to analyse the impact of various distance measuring factors on the quality of the outcomes of the clustering algorithm. Figure 3-2 indicates that for a total of 12 distance measurements, 15 datasets were used with four distance-based algorithms. All distance measurements in Figure 3-2 were examined, except for the weighted ED, which depends on the dataset and the clustering purpose. Figure 3-3 explains the design of this section briefly. All four distance-based methods were examined for each dataset, and the clustering quality of each method was evaluated based on the application of each of the 12 distance measurements, as illustrated in Figure 3-2. The experiment provides a total of 720 tests to evaluate the impact of distance measurements. Representation and comparison of this substantial number of

tests is a difficult task and could not be achieved using regular graphs and tables. Therefore, a unique illustration technique was created using heat mapped tables to display all the outcomes in a manner that they could be read and grasped rapidly. This technique is described in Section 4.2.

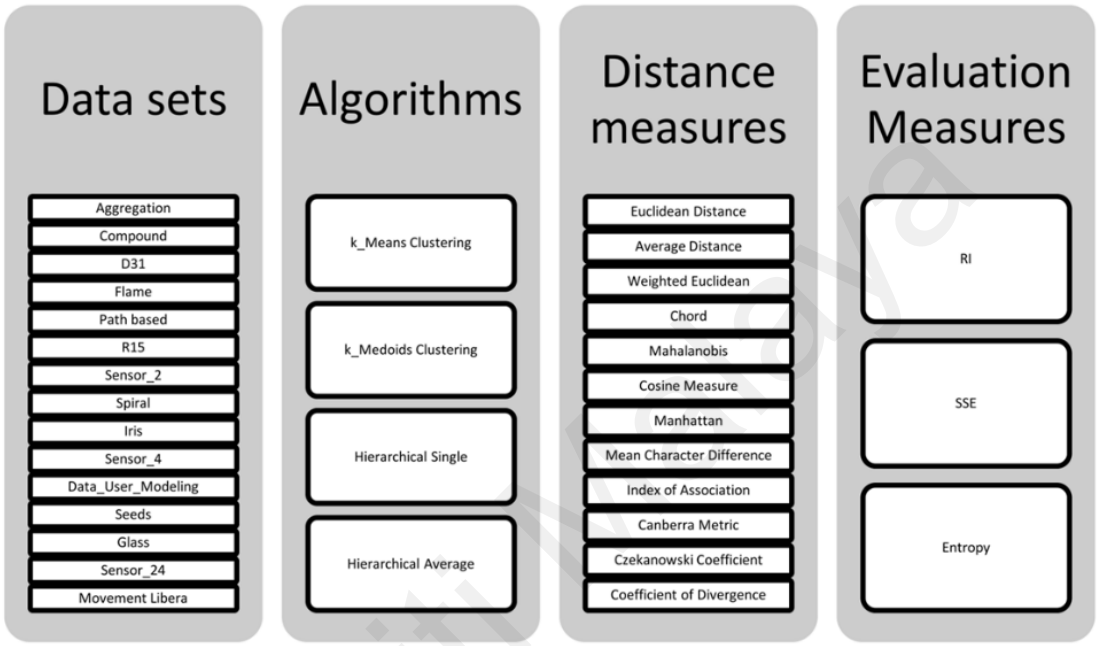

Figure 3-2: Process for evaluating distance measures

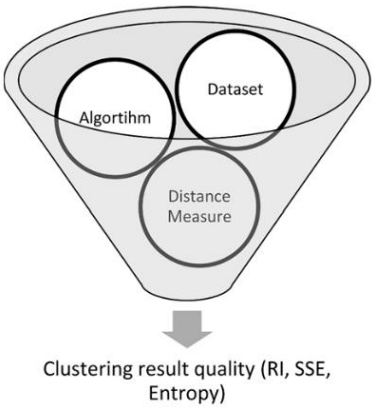

Figure 3-3: Distance measure evaluation components

(a) ***Evaluation of similarities measures***

This analysis utilises the RI to evaluate clustering outcomes resulting from different distance measurements. The RI is commonly used to measure the performance of clustering. It is a metric of agreement between two groups of items: the first set is generated by the clustering methods, and the other is identified by internal criteria.

(b) ***Analysis of variance (ANOVA) test***

An analysis of variance (ANOVA) test was used to demonstrate that distance measurements have a significant effect on the performance of clustering. For this reason, a null hypothesis was assumed: "distance measures have no critical impact on the performance of clustering". In the ANOVA test, if the *p*-value is low, it implies that there is little chance that the null hypothesis is true, and it can therefore be rejected. The ANOVA test, developed by Ronald Fisher (Fisher, 1992), analyses the differences between variables. It is a type of statistical test that indicates whether the mean of several individuals is equivalent, and it generalises the *t*-test for more than two groups. Furthermore, it is helpful for testing the means for statistical significance of more than two groups or variables. Statistical significance in statistics is accomplished when a *p*-value is less than the significance level (Cumming, 2011). The *p*-value is the possibility of achieving results that accept that the null hypothesis is true (Schlotzhauer, 2007).

3.2.4.2 Stage 2: Pre-processing

Before performing clustering analysis on pulses coming from the scintillator detector, a pre-processing phase must take place. The following figure presents the activities involved in the pre-processing phase.

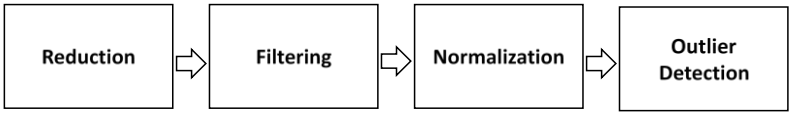

Figure 3-4- Pre-processing stages

3.2.4.3 Stage 3: The EFCA

The EFCA consists of four steps: pre-processing, epoch generation, aggregation, and assignment stages. Figure 3-5 provides a summary of the process in the EFCA:

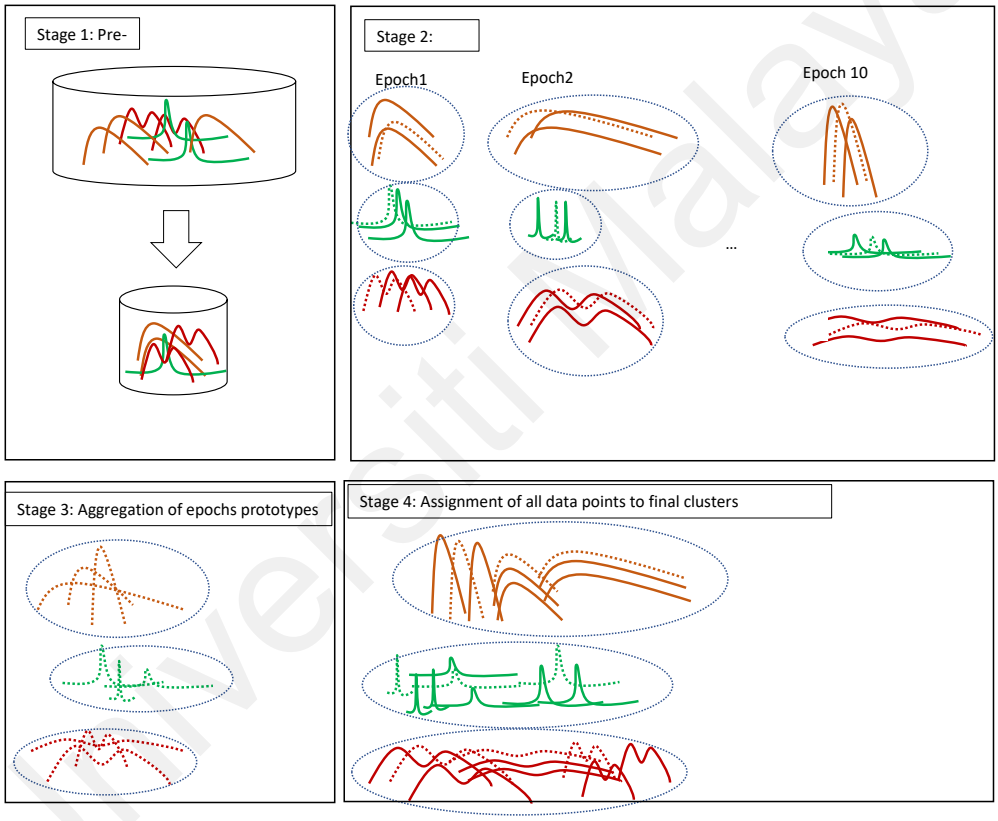

Figure 3-5: The EFCA clustering stages

Instant clustering approaches, regardless of the method, will cluster the whole dataset in a run and generate one set of cluster representation. In contrast, in evolving clustering, there are multiple epochs, and in each epoch, only a portion of data will be clustered, and a set of cluster representatives would exist for each epoch run. The EFCA attempts to cluster those data with more explicit clusters at early epochs, and data that are clustered

in final epochs can consequently be dismissed because of their obscurity. Clustering will be broken down to multiple epochs, and in each epoch, a membership matrix is used to cluster only those data points that have high similarity to prototypes, while the rest of the data points will be passed for further clustering in the next epoch.

(a) ***The motivation for epoch clustering***

The EFCA provides a mechanism that makes it possible to eliminate a portion of data that may contain noisy data, thereby leading to an improvement in the overall quality of clustering. This mechanism is based on two features, which are as follows:

1. The expectation in the EFCA is that the data clustered in earlier epochs have better quality, since they are clustered in the presence of more information (data), and the quality can decrease as data become sparse in upcoming epochs. This can be the case especially in a noisy dataset.
2. The EFCA provides an "epoch cut" – the option to disregard a few of the final epochs, which are estimated to contain noisy data.

(b) ***Motivation for the post-pruning approach***

To automatically identify and disregard obscure data, a heuristic approach, called "post-pruning", was introduced. The EFCA attempts to cluster data that are more explicit at early epochs, and later data that are clustered in final epochs can be disregarded because of their obscurity. The design of the EFCA clustering approach is illustrated in Figure 3-6.

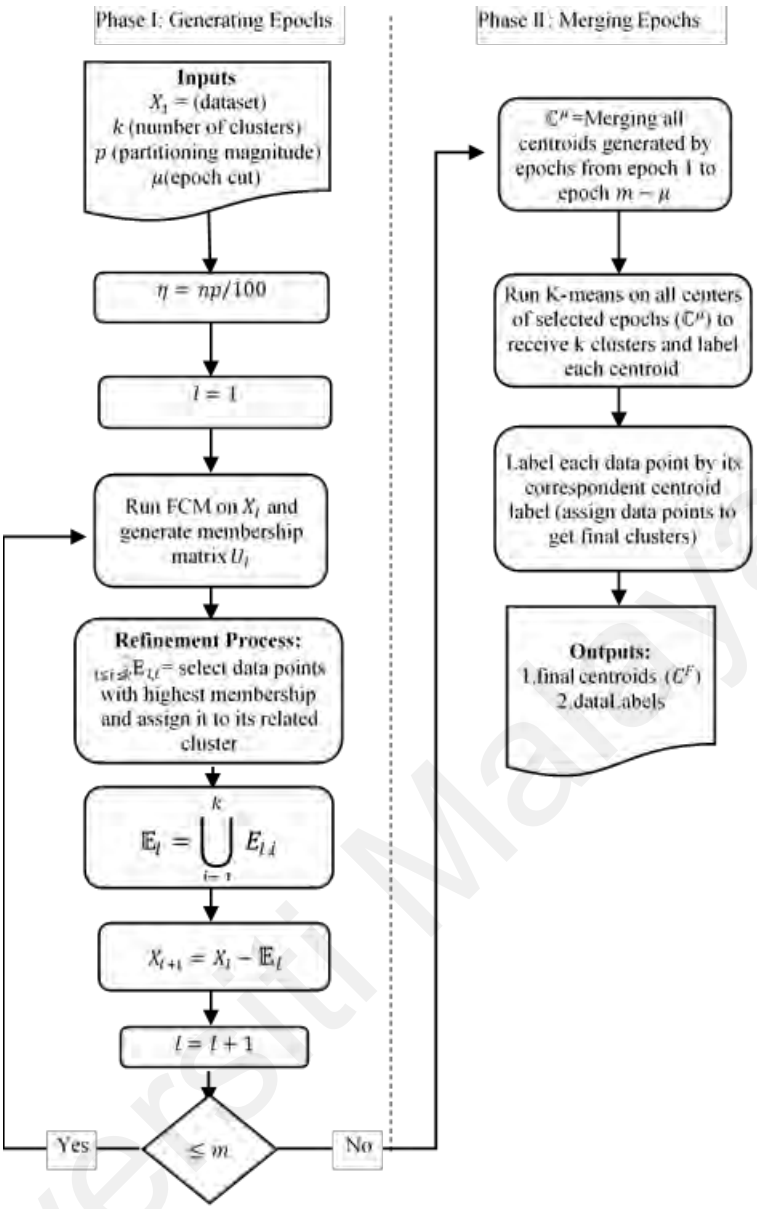

Figure 3-6: Design of the EFCA clustering approach

3.2.5 Analysis methods

After developing the proposed method, using the MATLAB software, all the steps were implemented. The EFCA was then applied to various datasets, and for each dataset, the proposed method was evaluated separately, then compared with standard methods, and finally analysed. Experimental details and results are discussed in CHAPTER 4:.

3.2.6 Evaluation method

A measure is needed to compare the quality of the EFCA with standard methods, *K*-means, and the FCM. In the presence of ground truth, an external clustering evaluation measure can be utilised. In this regard, the RI is a common performance metric for cluster assessment (Rand, 1971); it analyses the harmony between two partitions and demonstrates how near to ground truth the clustering outcomes are. It also compares EFCA cluster outcomes with conventional clustering methods.

The proposed model was evaluated experimentally to address the second objective of this study. Keogh and Kasetty (2003) conducted a noteworthy investigation of several time series research articles and concluded that the evaluation of clustering methods should comply with the previously mentioned recommendations, which are restated below:

- Verification of the method must be conducted on a different range of datasets (except when the method is only generated for a particular dataset).
- The dataset that is used must be readily available or published. A cautious design of studies must prevent bias in implementation.
- Wherever feasible, datasets and algorithms must be presented openly and publicly as free.
- A comparison of new similarity measurement methods must also be made with simple and reliable measurements, such as the ED.

The evaluation process is demonstrated in Figure 3-7.

Figure 3-7: Evaluation process of the proposed method

3.3 Chapter Summary

This chapter was intended to describe the methodology of this dissertation research. The method was chosen based on the problem statement and the research objectives. The chapter began with a research strategy, followed by a review of related work. Thereafter, the problem formulation was discussed, and the research objectives were defined.

In addition, it was clarified that to implement the proposed method and accomplish the objectives that were mentioned, a prerequisite is to indicate an accurate distance measure. Hence, the details of each step to determine the right distance measure were explained. The method proposed in this study is called the EFCA, which utilises the fuzzy membership concept to break down clustering into epochs, instead of running the clustering on all data at once. The motivation for using an EFCA's "epoch cut", which enables post-pruning ability, to eliminate obscure data points was elaborated: it results in higher clustering accuracy. The details of the proposed model and the techniques used in each step were then described. Finally, the evaluation strategy for the suggested approach was clarified in the last step of the research methodology framework. As mentioned, the RI utilises various types of datasets to evaluate the accuracy of the suggested approach. Further details about each component are explained in the next chapter.

CHAPTER 4:

IMPLEMENTATION AND EVALUATION

4.1 Introduction

Traditional methods of discriminating between neutron and gamma rays have been analogue-based; however, analogue-to-digital converters (ADCs) have recently opened the door to the digital world and provided the possibility to analyse the pulses coming from scintillators by digital approaches. Digital pulses can be treated as time series, and clustering is a method to group time series based on their similarity in shape. Unlike the TOF methodology, which is a common approach of discrimination, the clustering time series method requires no data source other than scintillator pulses.

The pulses can consequently and rapidly be discriminated into a neutron or gamma-ray, and even immediately after they come from the scintillator. This study aims to investigate different clustering methods in the discrimination task, it also presents a new approach, named the EFCA, and in this chapter, its implementation and the analysis are provided. The EFCA is tested against different clustering approaches and is analysed based on its quality of discrimination.

For convenience, the researcher divided this study into three phases. The rest of this chapter is devoted to describing in detail the procedure and evaluation of each of those phases in the research.

4.2 Phase 1: Investigating Accurate Distance Measures for Continuous Data

Before presenting the similarity measures for continuous data, a description of clustering should be presented. Provided that k is the number of clusters to be generated, the clustering is summarised as follows (Dunham, 2003):

Definition 1:

Given a dataset $D = \{v_1, v_2, \dots, v_n\}$ of data vectors and an integer value k, the clustering problem is defining a mapping $f: D \to \{1, \dots, k\}$, where each v_i is assigned to one cluster $C_j, 1 \leq j \leq k$. A cluster C_j contains precisely those data vectors mapped to it; that is, $C_j = \{v_i \mid f(t_i) = C_{j,}\ 1 \leq i \leq n, and\ v_i \in D\}$. Here, v_1, v_2 represent two data vectors defined as $v_1 = \{x_1, x_2, \dots, x_n\}, v_2 = \{y_1, y_2, \dots, y_n\}$, where x_i, y_i are called attributes.

4.2.1 Similarity or dissimilarity measures for continuous data

Similarity metrics for continuous data are discussed in this section. Some of these measures are often used for clustering reasons, whereas others have hardly been found in the literature.

4.2.1.1 *Minkowski*

Two specific cases of the Minkowski family include the ED and the Manhattan distance (Cha & Sung-Hyuk, 2007; Gan et al., 2007; J Han et al., 2006). The Minkowski distance is well-defined by $d_{min} = (\sum_{i=1}^{n} |x_i - y_i|^m)^{\frac{1}{m}},\ m \geq 1$, where m is a positive real number, and x_i and y_i are two vectors in an n-dimensional space. The Minkowski distance works well when the clusters are separated from one another or, in other words, when data are compacted inside the clusters; if the dataset does not meet this requirement, then the large-scale attributes would overtake the others (A. K. Jain, Murty, & Flynn, 1999; Mao & Jain, 1996). A problem the with Minkowski distance is that the largest attribute dominates the remainder of attributes. The solution to this issue is to normalise the continuous features (A. K. Jain, Murty, & Flynn, 1999).

A new version of the Minkowski measure has been suggested to overcome clustering barriers. Wilson and Martinez, for instance, provided a distance based

on nominal features and enhanced the Minkowski measure for continuous features (Wilson & Martinez, 1997).

4.2.1.2 *Manhattan distance*

The Manhattan distance is a particular type of Minkowski distance at m = 1. As with the Minkowski measure, the Manhattan distance is vulnerable to outliers. When used in clustering algorithms, the cluster shape would be hyper-rectangular (R. Xu & Wunsch, 2005). Research by Perlibakas demonstrated that one of the best distance measurements for PCA-based facial recognition is a modified model of this distance measure (Perlibakas, 2004). This measure is defined as $d_{man} = \sum_{i=1}^{n} |x_i - y_i|$.

4.2.1.3 *Euclidean distance*

The ED is likely the most recognised distance measure used for numerical data. This is a particular case of the Minkowski measure when m = 2. If the dataset contains isolated clusters or compact clusters, then the ED will perform well (A. K. Jain, Murty, & Flynn, 1999; Mao & Jain, 1996). However, although the ED is prevalent in clustering, it has a disadvantage: if two data points do not share feature characteristics, they may have a lower distance than the other pair of data points with the same feature values (A. K. Jain, Murty, & Flynn, 1999; Legendre & Legendre, 2012; Wang et al., 2002). A further concern with the ED as a Minkowski measurement family is that the data with the highest feature value would dominate the others. Using normalisation for the continuous feature can be the answer to such a problem (A. K. Jain, Murty, & Flynn, 1999).

4.2.1.4 *Average distance*

With regard to the ED disadvantage listed above, the average distance was introduced as a modified version to improve the results (Gan et al., 2007; Legendre & Legendre,

2012). For two data points x, y in an n-dimensional space, the average distance is defined as $d_{ave} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2\right)^{\frac{1}{2}}$.

4.2.1.5 *Weighted Euclidean distance*

If the corresponding importance is accessible for each feature, then the weighted ED – another modified version of the ED – can be used (Hand et al., 2001). This distance is defined as $d_{we} = (\sum_{i=1}^{n} w_i(x_i - y_i)^2)^{\frac{1}{2}}$, where w_i is the weight assigned to the ith component.

This distance measure is the only metric that was not used in this comparison because the calculation of weights is strongly linked to the dataset and the researcher's purpose for cluster analysis. This measure has been used to suggest a dynamic form for the fuzzy clustering algorithm. Ji et al.'s research is an example of using this measure (Ji et al., 2013).

4.2.1.6 *Chord distance*

Chord distance is another ED modification to tackle the deficiencies discussed above. It can also fix problems experienced by the measurement scale. Chord distance is described as the chord length that links two normalised pieces of data within a hypersphere with a radius value of 1. It is also possible to calculate this distance from un-normalised data (Gan et al., 2007). Chord distance is defined as $d_{chord} = \left(2 - 2\frac{\sum_{i=1}^{n} x_i y_i}{\|x\|_2 \|y\|_2}\right)^{\frac{1}{2}}$, where $\|x\|_2$ is the L^2-norm $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.

4.2.1.7 *Mahalanobis distance*

Unlike the Euclidean and Manhattan distances, which are independent of the dataset from which the two data points are driven, the Mahalanobis distance is a data-driven metric (Boriah, Chandola, & Kumar, 2008; Xu & Wunsch, 2005). A regularised Mahalanobis

distance may be used to extract hyper ellipsoidal clusters (Mao & Jain, 1996). On the other hand, the Mahalanobis distance can reduce distortions produced by the linear correlation between attributes with the implementation of a whitening transformation to the data or by using the square Mahalanobis distance (A. K. Jain, Murty, & Flynn, 1999). The Mahalanobis distance is defined by $d_{mah} = \sqrt{(x-y)\mathrm{S}^{-1}(x-y)^T}$, where *S* is the covariance matrix of the dataset (Gan et al., 2007; János Abonyi, 2007).

4.2.1.8 *Cosine measure*

The metric of cosine similarity is most often used in document similarity (J Han et al., 2006; R. Xu & Wunsch, 2005) and is defined as $Cosine(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\|x\|_2 \|y\|_2}$, where $\|y\|_2$ is the Euclidean norm of vector $y = (y_1, y_2, \ldots, y_n)$, defined as $\|y\|_2 = \sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}$. The cosine measure is invariant to rotation but is variant to linear transformations. It is also independent of vector length (Xu & Wunsch, 2005).

4.2.1.9 *Pearson correlation*

The Pearson correlation is extensively used in gene expression data clustering (Daxin Jiang et al., 2004; H. Wang et al., 2002; R. Xu & Wunsch, 2005). This metric of similarity assesses the similarity among the shapes of two patterns of gene expression. The Pearson correlation is defined by $Pearson(x,y) = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}}$, where μ_x and μ_y are the means for x and y respectively. The Pearson correlation has a shortcoming of being sensitive to outliers (Daxin Jiang et al., 2004; R. Xu & Wunsch, 2005).

The measures described above are the most frequently used for continuous data clustering. Table 4-1 presents a summary of these measures, along with some highlights of each.

Table 4-1: Similarity measures for continuous data (in time complexity; *n* is the number of dimensions of *x* and *y*)

Distance Measure	Equation	Time complexity	Advantages	Disadvantages	Applications
Euclidean Distance	$d_{euc} = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{\frac{1}{2}}$	*O*(*n*)	Common, easy to compute, and works well with datasets with compact or isolated clusters (Gan et al., 2007; A. K. Jain, Murty, & Flynn, 1999).	Sensitive to outliers (Gan et al., 2007; A. K. Jain, Murty, & Flynn, 1999).	*K*-means algorithm, fuzzy *C*-means algorithm (Ji et al., 2013).
Average Distance	$d_{ave} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2\right)^{\frac{1}{2}}$	*O*(*n*)	Better at handling outliers than the Euclidean distance (Legendre & Legendre, 2012).	Variables contribute independently to the measure of distance. Redundant values could dominate the similarity between data points (Hand et al., 2001).	*K*-means algorithm.
Weighted Euclidean	$d_{we} = \left(\sum_{i=1}^{n} w_i(x_i - y_i)^2\right)^{\frac{1}{2}}$	*O*(*n*)	The weight matrix allows for an increasing of the effect of more important data points than less important ones (Hand et al., 2001).	Same as the disadvantages of average distance.	Fuzzy *C*-means algorithm (Ji et al., 2013).
Chord	$d_{chord} = \left(2 - 2\frac{\sum_{i=1}^{n} x_i y_i}{\|x\|_2\|y\|_2}\right)^{\frac{1}{2}}$	*O*(3*n*)	Can work with un-normalised data (Gan et al., 2007).	It is not invariant to linear transformation (Xu & Wunsch, 2005).	Ecological resemblance detection (Legendre & Legendre, 2012).
Mahalanobis Distance	$\mathrm{d_{mah}} = \sqrt{(\mathrm{x}-\mathrm{y})\mathrm{S}^{-1}(\mathrm{x}-\mathrm{y})^{\mathrm{T}}}$	*O*(3*n*)	The Mahalanobis distance is a data-driven measure that can ease the distance distortion caused by a linear combination of attributes (Legendre & Legendre, 2012).	It can be expensive in terms of computation (Xu & Wunsch, 2005).	Hyper ellipsoidal clustering algorithm (Mao & Jain, 1996).
Cosine Measure	$\mathrm{Cosine}(\mathrm{x},\mathrm{y}) = \frac{\sum_{i=1}^{n} \mathrm{x_i y_i}}{\|\mathrm{x}\|_2\|\mathrm{y}\|_2}$	*O*(3*n*)	Independent of vector length and invariant to rotation (Xu & Wunsch, 2005).	It is not invariant to linear transformation (Xu & Wunsch, 2005).	It is mostly used in document similarity applications (Han et al., 2006; R. Xu & Wunsch, 2005).
Manhattan	$d_{man} = \sum_{i=1}^{n}(x_i - y_i)$	*O*(*n*)	Is common, and similarly to other Minkowski-driven distances, it works well with datasets with compact or isolated clusters (Gan et al., 2007).	Sensitive to outliers (Gan et al., 2007; Jain, Murty, & Flynn, 1999).	*K*-means algorithm.
Mean Character Difference	$d_{MCD} = \frac{1}{n}\sum_{i=1}^{n}\|x_i - y_i\|$	*O*(*n*)	*Results in accurate outcomes using the *K*-medoids algorithm.	*Low accuracy for high-dimensional datasets using *K*-means.	Partitioning and hierarchical clustering algorithms.
Index of Association	$d_{IOA} = \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{x_i}{\sum_{i=1}^{n} x_i} - \frac{y_i}{\sum_{i=1}^{n} y_i}\right\|$	*O*(3*n*)	-	*Low accuracy using *K*-means and *K*-medoids algorithms.	Partitioning and hierarchical clustering algorithms.
Canberra Metric	$d_{canb} = \sum_{i=1}^{n}\frac{\|x_i - y_i\|}{(x_i + y_i)}$	O(n)	*Results in accurate outcomes for high-dimensional datasets using the *K*-medoids algorithm.	-	Partitioning and hierarchical clustering algorithms.
Czekanowski Coefficient	$d_{czekan} = 1 - \frac{2\sum_{i=1}^{n} min(x_i, y_i)}{\sum_{i=1}^{n}(x_i + y_i)}$	*O*(2*n*)	*Results in accurate outcomes for medium-dimensional datasets using the *K*-means algorithm.	-	Partitioning and hierarchical clustering algorithms.
Coefficient of Divergence	$d_{canb} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - y_i}{x_i + y_i}\right)^2\right)^{\frac{1}{2}}$	*O*(*n*)	*Results in accurate outcomes using the *K*-means algorithm.	-	Partitioning and hierarchical clustering algorithms.
Pearson coefficient	$Pearson(x,y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}}$	*O*(2*n*)	*Results in accurate outcomes using the hierarchical single-link algorithm for high-dimensional datasets.	-	Partitioning and hierarchical clustering algorithms.
*Points marked by an asterisk are compiled based on this research's experimental results.					

Information regarding the datasets used for the accuracy assessment of distance measurements is listed in Table 4-2, In addition, the major outcomes are discussed in the next section.

Table 4-2: Dataset details

Dataset Name	Dimensions	Clusters	Vectors
Aggregation	2	7	788
Compound	2	6	399
D31	2	31	3100
Flame	2	2	240
Path based	2	3	300
R15	2	15	600
Sensor_2	2	4	5456
Spiral	2	3	312
Iris	4	3	150
Sensor_4	4	4	5456
Data_User_Modeling	5	4	258
Seeds	7	3	210
Glass	9	7	214
Sensor_24	24	4	5456
Movement Libera	90	15	360

4.2.2 Experiment

The studies were carried out through distance-based partitioning (*K*-means and *K*-medoids) and hierarchical methods. While various clustering measures exist, such as entropy, Jaccard, purity, and the SSE, the RI is likely the most widely used cluster validation measure (Aghabozorgi et al., 2015; Hubert & Arabie, 1985; Santos & Embrechts, 2009). Assuming that $S = \{o_1, o_2, \dots, o_n\}$ is a set of n elements and that two partitions of S are given to compare $C = \{c_1, c_2, \dots, c_r\}$, which is a partition of *S* into *r* subsets, and $G = \{g_1, g_2, \dots, g_s\}$, which is a partition of *S* into *s* subsets, the RI (R) is defined as follows:

Definition 2:

$$RI = \frac{a + b}{a + b + c + d} \quad (1)$$

where

- a is the number of pairs of vectors in S that are in the same set in C and in the same set in G.
- b is the number of pairs of elements in S that are in different sets in C and in different sets in G.
- c is the number of pairs of elements in S that are in the same set in C and in different sets in G.
- d is the number of pairs of elements in S that are in different sets in C and in the same set in G.

A revised version of the RI exists, namely, the ARI, which Hubert and Arabie have suggested (Hubert & Arabie, 1985) as a modification on the recognised RI's main problems. These problems occur when the expected value of the RI for two random partition does not hold a constant value (e.g., 0) or when the RI approached its upper vale (1) as the number of clusters increases. However, because our data sources do not have these problems, and since the outcomes obtained using the ARI followed the same pattern of RI outcomes, the RI was utilised for clustering validation in this research because of its success in the clustering industry.

Then, the values of the RI were normalised for the tests at this point. The normalised RI values are the interval of 0 and 1, and the following formula was used to address it:

$$z_i = \frac{r_i - \min(r)}{\max(r) - \min(r)}$$

where $r = (r_1, \dots, r_n)$ is the array of RIs resulted by each similarity measure.

4.2.3 Illustration technique

An overview of the outcomes of the normalised RI is shown in colour scale tables in Figure 4-1 and Figure **4-2**. Given this study's aim of examining and evaluating the accuracy of similarity measures for various dimensional datasets, the tables are arranged based on horizontally ascending dimensions of the dataset. Following the first column, which provides the names of the similarity measurements, the remaining table is split into two batches of columns (low- and high-dimensional), listing the normalised RIs for low- and high-dimensional datasets respectively. The final column, which is named "overall average" in this table, illustrates a general exploration of the most accurate measure of similarity. This form illustration design is used for all four algorithms.

In this analysis, the *K*-means and *K*-medoids algorithms were implemented as partitioning algorithms, and the RI was used for accuracy assessment. Since the outcomes of the *K*-means and *K*-medoids algorithms rely on the initial, randomly selected centres, and sometimes their accuracy could be affected by the local minimum trap, the test was repeated 100 times for each similarity measure. Thereafter, the maximum RI was used for comparison. An overview of the outcomes of the normalised RI is depicted in colour scale tables in Figure 4-1 and Figure **4-2**.

	Low Dimensional								High Dimensional						
Dimensions	2	2	2	2	2	2	2	2	4	4	5	7	9	24	90
	sensor_2	Aggregation	Compound	Flame	Pathbased	R15	Spiral	D31	Iris	sensor_4	User_Mod	Seeds	Glass	sensor_24	Libras moveme
Euclidean	0.881	0.934	0.982	1.000	0.932	1.000	0.880	0.999	0.546	0.000	1.000	0.972	0.999	0.488	0.505
Average	0.912	0.934	0.972	1.000	0.932	1.000	0.881	0.995	0.546	0.091	1.000	0.968	1.000	0.655	0.941
Cosine	0.708	0.518	0.580	0.273	0.027	0.887	1.000	0.918	1.000	0.186	0.158	0.987	0.988	0.596	0.364
Chord	0.708	0.520	0.580	0.273	0.027	0.886	1.000	0.918	1.000	0.186	0.161	0.987	0.996	0.707	0.941
Mahalanobis	0.891	0.929	1.000	0.977	1.000	1.000	0.883	1.000	0.546	0.050	0.909	0.972	0.976	0.000	0.677
Canberra	0.942	0.910	0.915	0.844	0.815	0.999	0.817	0.994	0.874	1.000	0.793	0.939	0.654	0.754	0.166
CoeffDiv	0.934	0.855	0.958	0.000	0.797	0.998	0.839	0.994	0.915	0.730	0.711	0.795	0.649	0.853	0.587
Czekan	1.000	1.000	0.906	0.823	0.899	0.998	0.851	1.000	0.794	0.941	0.826	1.000	0.981	0.572	0.000
IndOfAssoc	0.715	0.519	0.576	0.227	0.014	0.883	0.986	0.918	0.957	0.274	0.159	0.997	0.977	0.482	0.417
Manhattan	0.901	0.929	0.953	0.977	0.914	0.998	0.901	0.998	0.515	0.860	0.801	0.963	0.980	0.279	0.495
MCharDiff	0.901	0.929	0.948	0.977	0.914	0.998	0.898	0.999	0.515	0.870	0.789	0.963	0.974	0.509	0.254
Pearson	0.000	0.000	0.000	0.147	0.000	0.000	0.000	0.000	0.000	0.438	0.000	0.000	0.000	1.000	1.000

Figure 4-1: *K*-means colour scale table for normalised Rand index values (green represents the highest, and it changes to red, which is the lowest Rand index value)

	Low Dimensional								High Dimensional						
Dimensions	2	2	2	2	2	2	2	2	4	4	5	7	9	24	90
	sensor_2	Aggregation	Compound	Flame	Pathbased	R15	Spiral	D31	Iris	sensor_4	User_Mod	Seeds	Glass	sensor_24	Libras movemen
Euclidean	0.928	0.437	1.000	1.000	0.932	1.000	0.787	1.000	0.719	0.607	0.957	0.996	0.990	0.569	0.990
Average	0.825	0.972	0.967	1.000	0.932	1.000	0.785	0.996	0.719	0.649	0.928	0.996	0.992	0.543	1.000
Cosine	0.641	0.480	0.555	0.152	0.000	0.883	0.793	0.918	1.000	0.651	0.802	0.934	0.997	0.544	0.998
Chord	0.641	0.480	0.551	0.152	0.000	0.880	0.793	0.918	1.000	0.657	0.802	0.934	0.982	0.528	0.980
Mahalanobis	0.809	0.975	0.987	0.977	1.000	0.976	0.790	0.995	0.250	0.000	0.962	0.849	0.928	0.000	0.000
Canberra	0.996	0.909	0.918	0.783	0.916	0.998	1.000	0.992	0.794	1.000	0.000	0.991	0.734	1.000	0.941
CoeffDiv	1.000	0.825	0.918	0.000	0.929	0.975	0.786	0.988	0.915	0.784	0.000	0.796	0.604	0.716	0.982
Czekan	0.978	1.000	0.941	0.824	0.929	0.998	0.824	0.994	0.682	0.632	0.000	1.000	0.994	0.663	0.961
IndOfAssoc	0.674	0.468	0.568	0.255	0.409	0.884	0.743	0.919	0.915	0.408	0.784	0.978	0.991	0.635	0.987
Manhattan	0.947	0.949	0.979	0.932	0.903	0.999	0.806	0.995	0.546	0.477	0.977	0.981	0.994	0.501	0.976
MCharDiff	0.975	0.986	0.972	0.932	0.903	0.998	0.910	0.999	0.546	0.712	1.000	0.981	1.000	0.456	0.974
Pearson	0.000	0.000	0.000	0.143	0.196	0.000	0.000	0.000	0.000	0.006	0.813	0.000	0.000	0.400	0.990

Figure 4-2: *K*-medoids colour scale table for normalised Rand index values (green is the highest, and it changes colour to red, which is the lowest Rand index value)

The table for the ANOVA test with the framework is presented in Table 4-3, which displays all RI results for each dissimilarity measure and all pairs of datasets and methods. Each row reflects outcomes produced for a dataset and a clustering method with distance measurements.

Table 4-3: Rand index values used for ANOVA test (in the table HAverage stands for hierarchical average method, and HSingle stands for hierarchical single link)

Dataset	**Method**	**Distance or Similarity Measures**											
		Euclidean	**Average**	**Cosine**	**Chord**	**Mahalanobis**	**Canberra**	**CoeffDiv**	**Czekan**	**IndOfAssoc**	**Manhattan**	**MCharDiff**	**Pearson**
sensor_2	***k*-Means**	0.722	0.733	0.659	0.659	0.725	0.744	0.741	0.765	0.662	0.729	0.729	0.403
sensor_2	***k*-Medoids**	0.777	0.736	0.661	0.661	0.729	0.804	0.806	0.797	0.675	0.785	0.796	0.403
sensor_2	**HSingle**	0.432	0.432	0.355	0.355	0.432	0.432	0.432	0.431	0.365	0.432	0.432	0.405
sensor_2	**HAverage**	0.466	0.466	0.634	0.634	0.506	0.466	0.729	0.716	0.634	0.466	0.466	0.404
Aggregation	***k*-Means**	0.929	0.929	0.798	0.799	0.927	0.921	0.904	0.949	0.799	0.927	0.927	0.636
Aggregation	***k*-Medoids**	0.949	0.949	0.790	0.790	0.950	0.928	0.901	0.958	0.787	0.941	0.953	0.636
Aggregation	**HSingle**	0.926	0.926	0.574	0.574	0.926	0.619	0.927	0.927	0.550	0.926	0.926	0.635
Aggregation	**HAverage**	1.000	1.000	0.778	0.778	0.997	0.930	0.948	0.927	0.778	0.991	0.991	0.643
Compound	***k*-Means**	0.919	0.914	0.746	0.746	0.926	0.890	0.908	0.886	0.744	0.906	0.904	0.497
Compound	***k*-Medoids**	0.925	0.911	0.734	0.733	0.920	0.890	0.890	0.900	0.740	0.916	0.913	0.497
Compound	**HSingle**	0.890	0.890	0.415	0.415	0.896	0.895	0.898	0.891	0.415	0.712	0.712	0.497
Compound	**HAverage**	0.921	0.921	0.676	0.676	0.921	0.850	0.852	0.829	0.697	0.933	0.933	0.511
Flame	***k*-Means**	0.756	0.756	0.569	0.569	0.750	0.716	0.498	0.710	0.557	0.750	0.750	0.536
Flame	***k*-Medoids**	0.762	0.762	0.538	0.538	0.756	0.705	0.498	0.716	0.565	0.744	0.744	0.536

Flame	**HSingle**	0.541	0.541	0.522	0.522	0.541	0.531	0.531	0.541	0.522	0.541	0.541	0.538
Flame	**HAverage**	0.721	0.721	0.503	0.503	0.847	0.512	0.529	0.501	0.503	0.689	0.689	0.538
Pathbased	***k*-Means**	0.750	0.750	0.639	0.639	0.758	0.735	0.733	0.746	0.637	0.748	0.748	0.635
Pathbased	***k*-Medoids**	0.746	0.746	0.606	0.606	0.756	0.743	0.745	0.745	0.667	0.741	0.741	0.635
Pathbased	**HSingle**	0.338	0.338	0.362	0.362	0.340	0.339	0.338	0.338	0.362	0.338	0.338	0.635
Pathbased	**HAverage**	0.738	0.738	0.699	0.699	0.754	0.438	0.377	0.708	0.629	0.724	0.724	0.635
R15	***k*-Means**	0.999	0.999	0.949	0.948	0.999	0.999	0.998	0.998	0.947	0.998	0.998	0.552
R15	***k*-Medoids**	0.999	0.999	0.947	0.945	0.988	0.998	0.988	0.998	0.947	0.999	0.998	0.552
R15	**HSingle**	0.910	0.910	0.817	0.817	0.910	0.856	0.857	0.856	0.817	0.911	0.911	0.574
R15	**HAverage**	0.999	0.999	0.917	0.917	0.999	0.981	0.963	0.990	0.914	0.998	0.998	0.566
Spiral	***k*-Means**	0.554	0.554	0.562	0.562	0.555	0.550	0.552	0.553	0.562	0.556	0.556	0.496
Spiral	***k*-Medoids**	0.555	0.554	0.555	0.555	0.555	0.571	0.555	0.557	0.551	0.556	0.564	0.496
Spiral	**HSingle**	1.000	1.000	0.383	0.383	1.000	0.781	0.781	0.781	0.383	1.000	1.000	0.497
Spiral	**HAverage**	0.537	0.537	0.528	0.528	0.557	0.424	0.499	0.498	0.428	0.540	0.540	0.497
D31	***k*-Means**	0.994	0.992	0.956	0.956	0.995	0.992	0.992	0.994	0.956	0.994	0.994	0.528
D31	***k*-Medoids**	0.994	0.992	0.956	0.956	0.992	0.990	0.988	0.991	0.956	0.991	0.994	0.528
D31	**HSingle**	0.779	0.779	0.818	0.818	0.754	0.740	0.731	0.730	0.518	0.755	0.755	0.536
D31	**HAverage**	0.994	0.994	0.950	0.950	0.996	0.977	0.979	0.986	0.952	0.996	0.996	0.537
Iris	***k*-Means**	0.880	0.880	0.966	0.966	0.880	0.942	0.950	0.927	0.958	0.874	0.874	0.776
Iris	***k*-Medoids**	0.912	0.912	0.966	0.966	0.824	0.927	0.950	0.906	0.950	0.880	0.880	0.776
Iris	**HSingle**	0.777	0.777	0.772	0.772	0.343	0.753	0.753	0.772	0.772	0.776	0.776	0.772
Iris	**HAverage**	0.892	0.892	0.772	0.772	0.343	0.753	0.753	0.778	0.772	0.886	0.886	0.776
sensor_4	***k*-Means**	0.612	0.624	0.637	0.637	0.619	0.745	0.709	0.737	0.649	0.726	0.728	0.670
sensor_4	***k*-Medoids**	0.707	0.711	0.711	0.711	0.656	0.740	0.722	0.709	0.690	0.696	0.716	0.656
sensor_4	**HSingle**	0.341	0.341	0.345	0.345	0.346	0.451	0.339	0.333	0.345	0.338	0.338	0.651
sensor_4	**HAverage**	0.338	0.338	0.561	0.561	0.338	0.479	0.479	0.480	0.544	0.376	0.376	0.653
Data_User_Modeling	***k*-Means**	0.725	0.725	0.668	0.668	0.719	0.711	0.706	0.713	0.668	0.712	0.711	0.657
Data_User_Modeling	***k*-Medoids**	0.725	0.712	0.654	0.654	0.728	0.285	0.285	0.285	0.646	0.734	0.745	0.659
Data_User_Modeling	**HSingle**	0.309	0.309	0.301	0.301	0.304	0.302	0.302	0.305	0.302	0.299	0.299	0.311
Data_User_Modeling	**HAverage**	0.659	0.659	0.301	0.301	0.337	0.302	0.302	0.307	0.309	0.645	0.645	0.594
Seeds	***k*-Means**	0.876	0.874	0.884	0.884	0.876	0.859	0.782	0.891	0.890	0.872	0.872	0.359
Seeds	***k*-Medoids**	0.874	0.874	0.842	0.842	0.798	0.872	0.771	0.876	0.865	0.867	0.867	0.359
Seeds	**HSingle**	0.357	0.357	0.340	0.340	0.337	0.340	0.337	0.340	0.340	0.340	0.340	0.358
Seeds	**HAverage**	0.887	0.887	0.691	0.691	0.337	0.879	0.581	0.802	0.688	0.802	0.802	0.362
Glass	***k*-Means**	0.741	0.742	0.737	0.740	0.732	0.604	0.602	0.734	0.732	0.734	0.731	0.342
Glass	***k*-Medoids**	0.735	0.736	0.738	0.732	0.711	0.633	0.582	0.737	0.735	0.737	0.739	0.342
Glass	**HSingle**	0.304	0.304	0.308	0.308	0.309	0.293	0.294	0.308	0.308	0.308	0.308	0.342
Glass	**HAverage**	0.329	0.329	0.570	0.570	0.309	0.328	0.323	0.415	0.415	0.415	0.415	0.369
sensor_24	***k*-Means**	0.610	0.615	0.614	0.617	0.596	0.618	0.621	0.613	0.610	0.604	0.611	0.626
sensor_24	***k*-Medoids**	0.624	0.623	0.623	0.622	0.588	0.652	0.634	0.630	0.629	0.620	0.617	0.613
sensor_24	**HSingle**	0.347	0.347	0.346	0.346	0.353	0.346	0.347	0.346	0.346	0.345	0.345	0.349
sensor_24	**HAverage**	0.353	0.353	0.538	0.538	0.347	0.498	0.516	0.518	0.521	0.428	0.428	0.446

Libras movement	***k*-Means**	0.914	0.917	0.913	0.917	0.915	0.911	0.914	0.910	0.913	0.914	0.912	0.918
Libras movement	***k*-Medoids**	0.907	0.909	0.908	0.905	0.720	0.897	0.905	0.901	0.906	0.904	0.904	0.907
Libras movement	**HSingle**	0.187	0.187	0.202	0.202	0.131	0.183	0.183	0.187	0.192	0.187	0.187	0.296
Libras movement	**HAverage**	0.886	0.886	0.892	0.892	0.131	0.582	0.613	0.827	0.844	0.861	0.861	0.886

The ANOVA test result from the above table is presented in Figure 4-3.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	15647.2	11	1422.48	2.96	0.0008
Error	340550.9	708	481		
Total	356198.1	719			

Figure 4-3: ANOVA test result

The small *p*-value of 0.0008 indicates that differences between the means of the columns are significant. This result suggests that similarity assessments have a significant effect on the performance of clustering. Later in this analysis, an investigation is conducted to determine how these similarity measures affect the performance of the clustering.

4.2.4 Benchmarking similarity measures for partitioning methods

The outcomes of the *K*-means method are presented in Figure 4-1. The figure illustrates that, on the one hand, the Mahalanobis distance measure has the highest scores among all similarity measures for low-dimensional datasets. One the other hand, the divergence coefficient is the most accurate, with the largest RI values for high-dimensional data. Figure 4-2 contains outcomes for the *K*-medoids method. The mean character difference is the most accurate metric for low-dimensional datasets, whereas for high-dimensional datasets, the cosine metric reflects more accurate results. Overall, for most datasets, the mean character difference is highly accurate.

As an overall conclusion to the partitioning methods used in this research, average distance yields more accurate and reliable results for both methods. Moreover, it is the most accurate metric in the *K*-means method, and it also ranks second for the *K*-medoids method after mean character difference, with a small difference.

From another perspective, similarity measures in the *K*-means algorithm can be evaluated to determine which of them would yield a faster implementation of *K*-means. However, because of the likelihood of falling into the local minimum trap, the convergence of *K*-means and *K*-medoid algorithms is not guaranteed. Therefore, the algorithm was run 100 times to avoid bias towards this deficiency. Figure 4-4 displays two sample box charts developed using normalised data, representing the normalised iteration number required to converge each similarity measurement. Outcomes were gathered after running the *K*-means algorithm 100 times for each similarity measure and dataset.

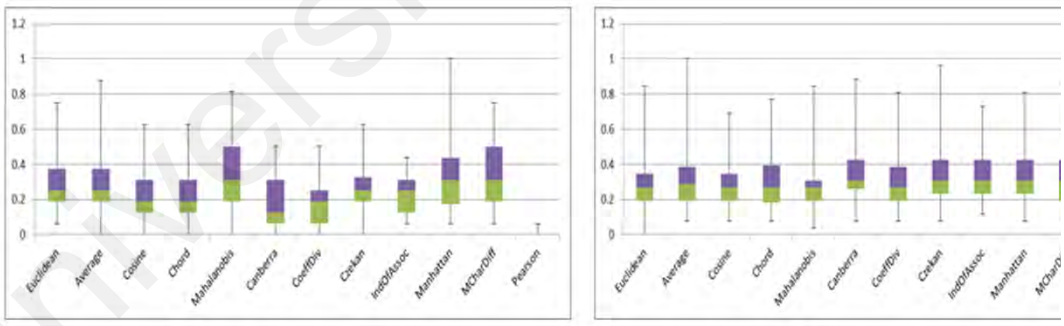

Figure 4-4: Sample box charts for *K*-means iteration counts created with a collection of normalised results after repeating the algorithm 100 times for each similarity measure and dataset

Figure 4-5 is a summarised colour scale table representing the mean and variance of iteration counts for all 100 algorithm runs. The Pearson measure has the fastest convergence in most datasets. After Pearson, the average and Euclidean distances are the fastest similarity measures in terms of convergence.

dimensions	2	2	2	2	2	2	2	2	4	4	5	7	9	24
	sensor_2	Aggregation	Compound	Flame	Pathbased	R15	Spiral	D31	Iris	sensor_4	Data_User_Modeling	Seeds	Glass	sensor_24
Euclidean	0.1682 (0.0686)	0.2040 (0.1007)	0.1728 (0.0969)	0.3277 (0.1516)	0.1728 (0.0969)	0.1094 (0.045)	0.3349 (0.1932)	0.1317 (0.0445)	0.3037 (0.1667)	0.1267 (0.0646)	0.1434 (0.1034)	0.1804 (0.0688)	0.2900 (0.1254)	0.2199 (0.1616)
Average	0.1658 (0.0689)	0.1922 (0.1011)	0.1643 (0.0806)	0.3504 (0.1594)	0.1643 (0.0806)	0.1073 (0.0391)	0.3558 (0.2448)	0.1349 (0.0503)	0.3018 (0.181)	0.1354 (0.0843)	0.1462 (0.0926)	0.1784 (0.0749)	0.2783 (0.1188)	0.2159 (0.1541)
Cosine	0.3265 (0.1725)	0.4033 (0.1966)	0.5253 (0.2831)	0.4179 (0.1491)	0.5253 (0.2831)	0.3191 (0.1617)	0.3166 (0.1904)	0.4018 (0.1468)	0.2254 (0.1191)	0.2176 (0.1216)	0.1104 (0.0695)	0.3076 (0.2268)	0.3150 (0.1455)	0.1762 (0.116)
Chord	0.2912 (0.1503)	0.3701 (0.2014)	0.5373 (0.2953)	0.4331 (0.1471)	0.5373 (0.2953)	0.2773 (0.1246)	0.2973 (0.1858)	0.3827 (0.1792)	0.2298 (0.1259)	0.2337 (0.1294)	0.1173 (0.0818)	0.3472 (0.2752)	0.3320 (0.1776)	0.1856 (0.1116)
Mahalanobis	0.1737 (0.0649)	0.2073 (0.0997)	0.1775 (0.0771)	0.3769 (0.1769)	0.1775 (0.0771)	0.1068 (0.0414)	0.2955 (0.168)	0.1333 (0.0482)	0.3415 (0.1849)	0.1242 (0.0745)	0.1636 (0.0958)	0.1867 (0.0727)	0.2741 (0.1265)	0.2034 (0.1173)
Canberra	0.2099 (0.0883)	0.2934 (0.1412)	0.2192 (0.0932)	0.4792 (0.1942)	0.2192 (0.0932)	0.1145 (0.0586)	0.2704 (0.1349)	0.1835 (0.068)	0.1824 (0.1349)	0.2236 (0.1733)	0.2243 (0.1569)	0.2124 (0.0819)	0.1486 (0.0704)	0.2873 (0.1987)
CoeffDiv	0.1965 (0.096)	0.2813 (0.1306)	0.2314 (0.0893)	0.4154 (0.1508)	0.2314 (0.0893)	0.1076 (0.0511)	0.2546 (0.1034)	0.1660 (0.0555)	0.1730 (0.1123)	0.1840 (0.0949)	0.1481 (0.0957)	0.2329 (0.0988)	0.1459 (0.1025)	0.3643 (0.2578)
Czekan	0.1976 (0.0787)	0.2755 (0.1197)	0.1780 (0.0808)	0.3592 (0.1417)	0.1780 (0.0808)	0.1269 (0.061)	0.2326 (0.0889)	0.1446 (0.0649)	0.2639 (0.1334)	0.1806 (0.1043)	0.1730 (0.1094)	0.1639 (0.0607)	0.2850 (0.1323)	0.1999 (0.1484)
IndOfAssoc	0.4238 (0.2591)	0.3800 (0.1871)	0.4116 (0.2462)	0.5385 (0.1706)	0.4116 (0.2462)	0.2824 (0.1639)	0.3177 (0.1794)	0.3686 (0.1423)	0.2254 (0.0988)	0.1636 (0.0745)	0.1480 (0.0953)	0.3033 (0.2167)	0.2640 (0.1029)	0.1893 (0.1007)
Manhattan	0.2578 (0.1074)	0.1807 (0.0654)	0.1748 (0.0721)	0.4040 (0.1764)	0.1748 (0.0721)	0.1310 (0.0617)	0.2477 (0.1458)	0.1261 (0.0489)	0.3371 (0.2081)	0.1857 (0.0798)	0.1365 (0.0807)	0.1830 (0.0755)	0.2682 (0.1116)	0.2109 (0.1283)
MCharDiff	0.2841 (0.1307)	0.1966 (0.0767)	0.1676 (0.0776)	0.4003 (0.2006)	0.1676 (0.0776)	0.1231 (0.055)	0.2323 (0.1332)	0.1251 (0.0519)	0.3333 (0.1874)	0.1828 (0.0797)	0.1496 (0.088)	0.1997 (0.076)	0.2693 (0.1355)	0.1943 (0.1288)
Pearson	0.0000 (0)	0.0000 (0)	0.0000 (0)	0.0000 (0)	0.0000 (0)	0.0000 (0)	0.0023 (0.0068)	0.0000 (0)	0.0006 (0.0062)	0.0255 (0.0256)	0.1272 (0.094)	0.0035 (0.0093)	0.0112 (0.0159)	0.1886 (0.1052)

Figure 4-5: Colour scale table for iteration count mean and variance (green is the lowest, and it changes colour to red, which represents the greatest iteration count value)

With regard to the RI and iteration rate, the average measure is demonstrated to be not only accurate in most datasets for both *K*-means and *K*-medoids algorithms, but also the second-fastest performance similarity measure after Pearson, rendering it a secure option when clustering is needed using *K*-means or *K*-medoids algorithms.

4.2.5 Benchmarking similarity measures for hierarchical methods

The impact of different similarity measures on *K*-means and *K*-medoids algorithms as partitioning algorithms was assessed and contrasted in the previous section. The findings for the single-link and group average algorithms, which are two hierarchical clustering algorithms, are now discussed in terms of the RI for each similarity measure in this section. Figure 4-6 and Figure 4-7 present the outcomes in sample bar charts, which include six sample datasets. Since bar charts would be jumbled for all datasets and similarity assessments, the findings are presented using colour scale tables for easier comprehension and discussion. As addressed in the last section, Figure 4-8 and Figure

4-9 are two colour scale tables depicting the normalised RI values for each similarity measure. The results in Figure 4-8 for the single-link algorithm illustrate that for low-dimensional datasets, the Mahalanobis distance is the most accurate similarity measure, and Pearson is the best among other measures for high-dimensional datasets. Furthermore, the overall average column in this figure indicates that most of the time, Pearson presents the highest accuracy, and the average and Euclidean distances are among the most accurate measures. For the group average algorithm, as seen in Figure 4-9, the Euclidean and average distances are the best among all similarity measures for low-dimensional datasets, whereas for high-dimensional datasets, cosine and chord are the most accurate measures. In the group average algorithm, the Manhattan distance and the mean character difference generally have the best overall RI results, followed by the ED and average distance measures. Considering the overall results, it is clear that the average measure is always among the best measurements, and it is best for both single-link and group average algorithms.

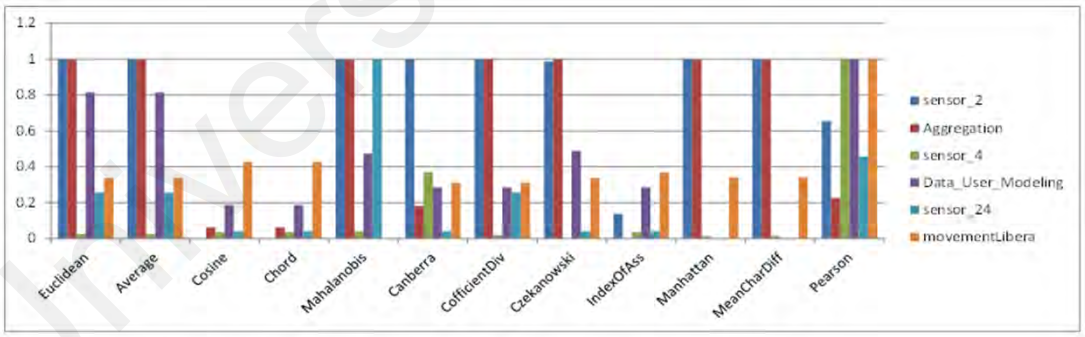

Figure 4-6: Bar chart of normalised Rand index values for selected datasets using the single-link algorithm

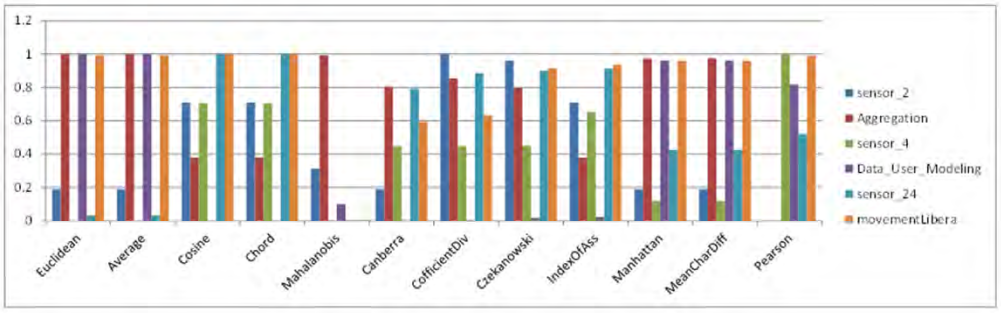

Figure 4-7: Bar chart of normalised Rand index values for selected datasets using the group average algorithm

	Low Dimensional								High Dimensional							
Dimension	2	2	2	2	2	2	2	2	4	4	5	7	9	24	90	Overall Average
	sensor_2	Aggregation	Compound	Flame	Pathbased	R15	Spiral	D31	Iris	sensor_4	User_Mod	Seeds	Glass	sensor_24	Libras movement	
Euclidean	1.000	0.996	0.984	0.000	0.996	0.194	1.000	0.869	1.000	0.025	0.814	0.939	0.212	0.257	0.337	0.642
Average	1.000	0.996	0.984	0.000	0.996	0.194	1.000	0.869	1.000	0.025	0.814	0.939	0.212	0.257	0.337	0.642
Cosine	0.000	0.063	0.000	0.080	0.722	0.007	0.000	1.000	0.989	0.036	0.186	0.160	0.298	0.041	0.428	0.267
Chord	0.000	0.063	0.000	0.080	0.722	0.007	0.000	1.000	0.989	0.036	0.186	0.160	0.298	0.041	0.428	0.267
Mahalanobis	1.000	0.998	0.995	0.008	0.996	1.000	1.000	0.786	0.000	0.040	0.474	0.000	0.326	1.000	0.000	0.575
Canberra	1.000	0.182	0.992	0.004	0.837	0.007	0.519	0.739	0.945	0.370	0.285	0.151	0.000	0.041	0.313	0.426
CoeffDiv	1.000	1.000	1.000	0.000	0.840	0.007	0.519	0.709	0.945	0.019	0.285	0.004	0.011	0.257	0.313	0.461
Czekan	0.988	1.000	0.984	0.000	0.837	0.007	1.000	0.705	0.989	0.000	0.489	0.151	0.298	0.041	0.338	0.522
IndOfAssoc	0.139	0.000	0.000	0.080	0.722	0.007	0.000	0.000	0.989	0.036	0.285	0.151	0.298	0.041	0.369	0.208
Manhattan	1.000	0.996	0.614	0.000	1.000	0.000	1.000	0.789	0.999	0.016	0.000	0.151	0.298	0.000	0.340	0.480
MCharDiff	1.000	0.996	0.614	0.000	1.000	0.000	1.000	0.789	0.999	0.016	0.000	0.151	0.298	0.000	0.340	0.480
Pearson	0.654	0.225	0.169	1.000	0.000	0.601	0.874	0.057	0.989	1.000	1.000	1.000	1.000	0.458	1.000	0.669

Figure 4-8: Colour scale table of normalised Rand index values for the single-link method (green is the highest, and it changes colour to red, which represents the lowest Rand index value)

	Low Dimensional								High Dimensional							
Dimensions	2	2	2	2	2	2	2	2	4	4	5	7	9	24	90	Overall Av
	sensor_2	Aggregation	Compound	Flame	Pathbased	R15	Spiral	D31	Iris	sensor_4	User_Mod	Seeds	Glass	sensor_24	Libras movement	
Euclidean	0.190	1.000	0.972	0.958	1.000	1.000	0.636	0.996	1.000	0.000	1.000	1.000	0.079	0.032	0.992	0.724
Average	0.190	1.000	0.972	0.958	1.000	1.000	0.636	0.996	1.000	0.000	1.000	1.000	0.079	0.032	0.992	0.724
Cosine	0.709	0.379	0.391	0.856	0.811	0.539	0.004	0.901	0.781	0.706	0.000	0.645	1.000	1.000	1.000	0.648
Chord	0.709	0.379	0.391	0.856	0.811	0.539	0.004	0.901	0.781	0.706	0.000	0.645	1.000	1.000	1.000	0.648
Mahalanobis	0.313	0.992	0.973	1.000	1.000	0.000	1.000	1.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.425
Canberra	0.190	0.804	0.803	0.163	0.959	0.495	0.030	0.959	0.746	0.448	0.003	0.987	0.073	0.791	0.593	0.536
CoeffDiv	1.000	0.853	0.808	0.000	0.917	0.460	0.081	0.963	0.746	0.448	0.003	0.444	0.056	0.885	0.633	0.553
Czekan	0.961	0.797	0.754	0.879	0.981	0.644	0.000	0.978	0.791	0.450	0.018	0.846	0.407	0.899	0.914	0.688
IndOfAssoc	0.709	0.378	0.440	0.669	0.803	0.780	0.004	0.904	0.781	0.652	0.022	0.640	0.407	0.914	0.937	0.603
Manhattan	0.190	0.974	1.000	0.922	0.999	0.644	0.543	0.999	0.988	0.119	0.962	0.846	0.407	0.425	0.960	0.732
MCharDiff	0.190	0.974	1.000	0.922	0.999	0.644	0.543	0.999	0.988	0.119	0.962	0.846	0.407	0.425	0.960	0.732
Pearson	0.000	0.000	0.000	0.685	0.000	0.869	0.107	0.000	0.789	1.000	0.818	0.047	0.230	0.521	0.992	0.404

Figure 4-9: Colour scale table of normalised Rand index values for group average (green is the highest, and it changes colour to red, which signifies the lowest Rand index value)

An overview of the outcomes on the *K*-means, *K*-medoids, single-link, and group average algorithms suggests that the average and Euclidean measures are frequently among the most accurate measures for all four algorithms. On the one hand, Figure 4-10 illustrates the overall average RI of all four algorithms, and all 15 datasets also uphold the same conclusion. Figure 4-11, on the other hand, presents the average RI for the four

algorithms separately. It can be inferred that the average and Euclidean measures are more accurate than other measures.

Furthermore, by using the *K*-means algorithm, this similarity measure is the fastest after Pearson in terms of convergence.

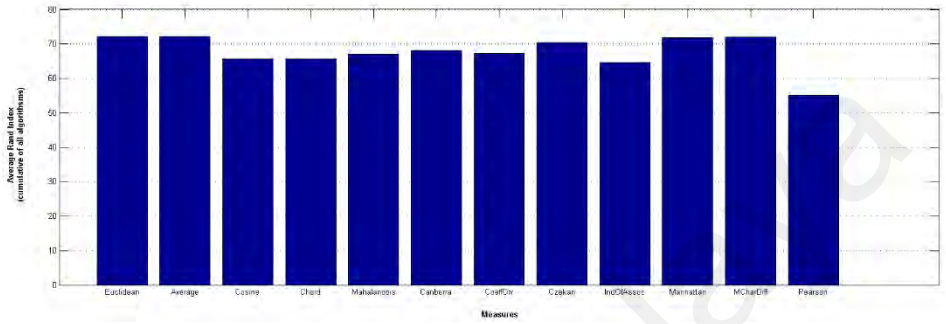

Figure 4-10: Overall RI average

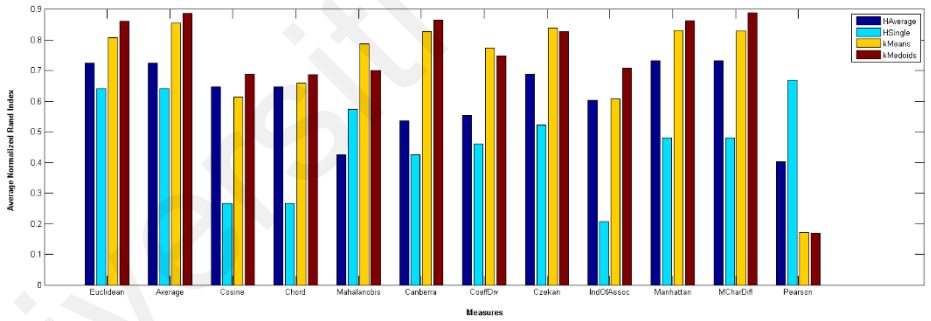

Figure 4-11: Average RI for four algorithms

4.2.6 Concluding remarks

Selecting the correct distance measure is one of the problems faced by experts and researchers when trying to implement a distance-based clustering algorithm in a dataset. The variety of similarity assessments can cause confusion and trouble in selecting an appropriate measurement. Moreover, these assessments can be performed differently for datasets with different dimensionalities. This research attempted to explain which

similarity measurements are more suitable for low-dimensional versus high-dimensional datasets. In this experiment, similarity measurements for numerical clustering data in distance-based algorithms were compared and benchmarked using 15 datasets classified as low- and high-dimensional datasets. The accuracy evaluation based on the RI was then studied for each similarity measurement, and the most appropriate similarity measurements were addressed for each of the low- and high-dimensional datasets for four well-known distance-based algorithms. Overall, the findings indicate that for all clustering algorithms used in this research, the average distance and ED are among the most accurate measures. Furthermore, when *K*-means is the target clustering algorithm, this measure is one of the fastest in terms of convergence.

4.3 Phase 2: Pre-processing

This study used a neutron and gamma-ray dataset collected by Savran et al. (2010). The pre-processing steps are briefly explained next.

4.3.1 Reduction

The first step is to determining the baseline of the pulse, which must be subtracted from the pulse. For the duration of pre-triggered, a constant baseline was defined to be10-30 ns before the real pulse in this test (Savran et al., 2010).

4.3.2 Filtering

Using the DFT, a finite impulse response (FIR) filter [0.25, 0.25, 0.25, 0.25, 0.25],demonstrated in Equation 4-1, is implemented twice for the signals to decrease the impact of high-frequency noise signals (Savran et al., 2010).

$$y(n) = \sum_{k=0}^{M-1} h(k)x(n-k) \qquad 4\text{-}1$$

4.3.3 Normalisation

To prevent dependency of clustering on the height of the pulses, and as the ED is utilised in the clustering process as a distance measure, normalising signals before the clustering procedure is necessary. The following formula was used for normalising the pulses between 0 and 1:

$if\ Pulse = p_1, p_2, \dots, p_n$

$$Normalized\ (p_i) = \frac{p_i - P_{min}}{P_{max} - P_{min}} \qquad 4\text{-}2$$

where $P_{min} = the\ minimum\ value\ for\ pulse\ P,\ P_{max} = the\ maximum\ value\ for\ pulse\ P$

4.3.4 Outlier detection

Last but not least is the outlier detection process. Pulses that are created or deformed by different types of noises, which are caused either by digital devices or when passing the wire, should be eliminated. To eliminate outliers, an average of the pulses was calculated, and then a maximum distance of 3 from the average was chosen as the border for accepted and outlier data. This means that pulses with a distance more than 3 from the average were flagged as outliers and eliminated from the dataset. In Figure 4-12, the distance of outliers is illustrated in chart (a), and the outlier pulse is depicted in chart (b).

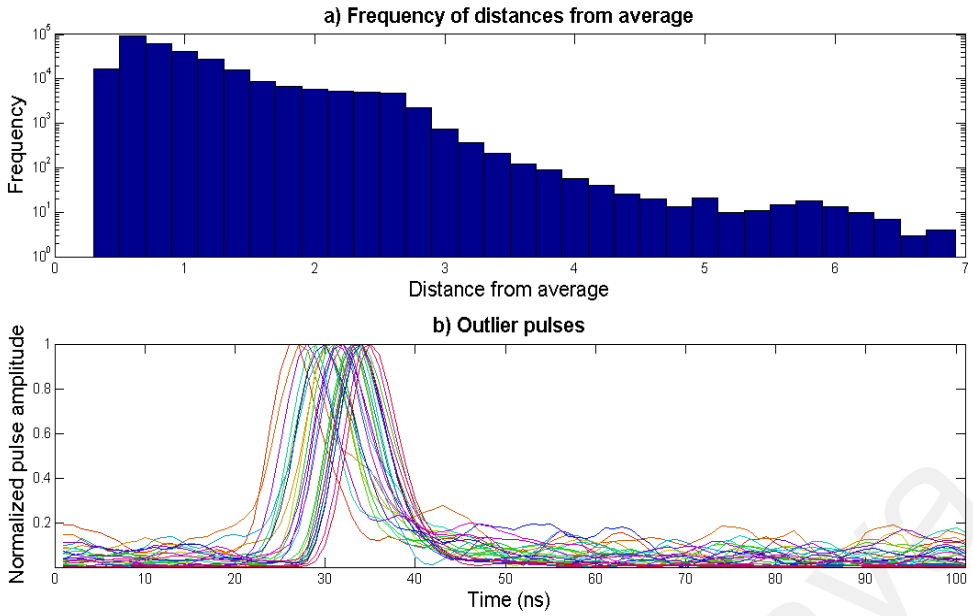

Figure 4-12: Pre-processing step for outlier detection

4.4 Phase 3: Evolving fuzzy clustering approach (EFCA)

In recent years, digital approaches have attracted researchers who are working on particle discrimination. Ronchi and colleagues (Ronchi et al., 2009) and another study by Akkoyun (Akkoyun, 2013) have demonstrated that ANNs can correctly classify gamma-ray and neutron pulses. Furthermore, Savran and his colleagues (2010) recently demonstrated that the FCM clustering algorithm is well suited for discrimination purposes.

This study is an attempt to introduce a new clustering method that can improve the accuracy of the clustering approach. To demonstrate that this new clustering approach is a general approach that can also be applied in other datasets, the method was evaluated not only on a neutron and gamma-ray dataset but also in a number of other multivariate and time series datasets.

Let X be defined as a dataset that can be shown as $\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{n}\}$, where n is the number of instances in a dataset, $x_i \in \mathbb{R}^n$, k is the number of clusters, and $k \in \mathbb{N}$. Furthermore, p is defined as the partitioning magnitude, which clarifies how much of the

data should be clustered in each epoch; p is a value between 1 and 100, which divides 100 ($p \mid 100$), and m number of epochs will be calculated with the following equation:

$$m = \frac{100}{p}, \quad \begin{cases} 0 < p \leq 100, \\ s.t. \;\; p \mid 100 \end{cases} \tag{4-3}$$

and η is the number of data points in each epoch that can be obtained as follows:

$$\eta = [\![\frac{pn}{100}]\!] \tag{4-4}$$

where η is the round number to the nearest integer to $\frac{pn}{100}$. In those cases where $\frac{pn}{100}$ is not an integer, the final epoch will receive excessive or less data points.

4.4.1 Epoch operation

In each epoch, fuzzy clustering is run on $X_l, \quad l = \{1, \dots, m\}$, where l is the number of epochs. The outcome is the membership matrix U_l, so that each element u_{ij} of this matrix expresses the membership value of data point j to cluster i, which can be written as follows:

$$FCM(X_l) \rightarrow U_l = \begin{pmatrix} u_{1,1} & \cdots & u_{1,\varphi_l} \\ \vdots & \ddots & \vdots \\ u_{k,1} & \cdots & u_{k,\varphi_l} \end{pmatrix} \tag{4-5}$$

where k is the number of clusters, and $\varphi_l = n - (l-1)\eta$ is the number of data points that exist in the input dataset of the l^{th} epoch.

Refinement Process (RP) in epoch l: In the next step, all elements of membership matrix U_l are sorted from maximum to minimum. Furthermore, $d_{l,q}$, $\{1 \leq q \leq k\varphi_l\}$ is defined as the element of U_l such that $d_{l,1}$ is the first element with the highest membership value in the U_l membership matrix, so $d_{l,2}$ is the second element with the highest membership value in the membership matrix and so on, with the last element $d_{l,k\varphi}$ being

the minimum of all elements in the above matrix, where the $k\varphi_l = k \times \varphi_l$. In other words,

$$D_l = \left\{ d_{l,q} \mid d_{l,q} = u_{ij}^{q} \geq d_{l,q+1} = u_{ij}^{q+1}, 1 \leq i \leq k, 1 \leq j \leq \varphi_l \right\} = \left\{ d_{l,1}, d_{l,2}, \dots, d_{l,k\varphi} \right\}, \quad 1 \leq q \leq k\varphi_l \tag{4-6}$$

Therefore, all elements in U_l are sorted out in epoch l from maximum to minimum as follows:

$$d_{l,1} \geq d_{l,2} \geq \cdots \geq d_{l,\varphi_l} \tag{4-7}$$

The next step is to select $d_{l,1}$ to $d_{l,\eta}$ from the above sorted array and assign their related data points to the respective clusters. The rest of the data points are neglected and passed to the dataset for the next epoch.

Each epoch has three outputs, which are as follows:

1. k clusters and their respective data points;
2. k centroids for each cluster; and
3. the remaining dataset for next epoch.

This process is repeated for m (number of epochs) times.

4.4.2 Unifying the epoch centres

After the epochs are finalised, two sets would be yielded as the output of the epochs:

1. The first output is $\mathbb{C}$, which is the set of cluster centers that are generated in all epochs. In other words,

$$\mathbb{C} = \bigcup_{l=1}^{m} C_l \tag{4-8}$$

where m is the number of epochs, such that C_l is the set of all centroids that are generated in the l^{th} epoch, so that

$$C_l = \{c_{l,1}, c_{l,2}, \dots, c_{l,i}, \dots, c_{l,k}\}, \quad 1 < i < k \tag{4-9}$$

Therefore, $\mathbb{C}$ can also be written as follows:

$$\mathbb{C} = \{c_{1,1}, \dots, c_{1,k}, c_{2,1}, \dots, c_{2,k}, \dots, c_{l,1}, \dots, c_{l,k}\} \tag{4-10}$$

where k is the number of clusters. Now each $c_{l,i}$ is the centroid of cluster i generated in epoch l.

2. The second output is $\overline{\mathbb{E}}$, which is the set of datasets clustered in all epochs. In other words,

$$\overline{\mathbb{E}} = \bigcup_{l=1}^{m} \mathbb{E}_l \tag{4-11}$$

such that $\mathbb{E}_l$ is a set that is the union of all clustered data points that have been clustered after a refinement process in epoch l; that is,

$$\mathbb{E}_l = \bigcup_{i=1}^{k} E_{l,i} \tag{4-12}$$

such that $E_{l,i}$ is a set of data points in epoch l that are assigned to the cluster with centroid $c_{l,i}$.

A one-to-one relationship exists between the set C_l and $\mathbb{E}_l$, which means that for each centroid in C_l, there is a dataset in $\mathbb{E}_l$.

To obtain the final result, which is the clustering of main dataset into k clusters, a final round of clustering using the k-means method is executed on $\mathbb{C}$. This results in k clusters of centroids that can be demonstrated by C_i^F, so that $\bigcup_{i=1}^{k} C_i^F$ (F indicates the final round

of clustering) is the set of all centroids generated by all epochs. Then, the data points related to each centroid are assigned to their respective cluster. This process results in k clusters of data points that were intended to obtain. These final clusters can be demonstrated by $E_i^F, 1 \leq i \leq k$.

4.4.3 Epoch cut

The EFCA provides a mechanism that makes it possible to eliminate a portion of data that may contain noisy data, resulting in an improvement in the overall quality of clustering. This mechanism is based on the following two features:

1) The expectation in the EFCA is that the data clustered in earlier epochs have better quality, since they are clustered in the existence of more data, and the quality can be decreased as data become sparse in upcoming epochs, especially when the dataset contains noisy data.
2) The EFCA provides an "epoch cut", which provides the option to disregard final epochs that may contain noisy data.

The epoch cut is μ, where $0 \leq \mu < m$ is the number of epochs that are going to be eliminated from the final clustering step. This means that set $\mathbb{C}$ should be replaced by set $\mathbb{C}^{\mu}$ as defined below:

$$\mathbb{C}^{\mu} = \bigcup_{l=1}^{m-\mu} C_l \qquad 4\text{-}13$$

Therefore, in the final step, k-mean clustering takes place on $\mathbb{C}^{\mu}$ instead of $\mathbb{C}$.

The above procedure is demonstrated in the pseudocode provided in Algorithm I.

Algorithm I: EFCA Pseudocode

Procedure EFCA (X, k, p, μ) ►k:number of clusters, p:partitioning magnitude, μ:epoch cut
  $m \leftarrow 100/p$
  $n \leftarrow length(X)$
  $\eta \leftarrow round((pn)/100)$
  $l \leftarrow 1$
  $X[l] \leftarrow X$
  While $l \leq m$ ***do*:** ►m is the number of epochs
    $[U[l], Centers] \leftarrow FCM(X, k)$
    $C[l] \leftarrow Centers$
    $D[l] \leftarrow$ top η maximum membership values in $U[l]$
    $\mathbb{E}[l] \leftarrow$ respective data points related to membership values in D
    $X[l+1] = X[l] - \mathbb{E}[l]$ ►Eliminating clustered data points
    $l \leftarrow l + 1$
  end while
  $\mathbb{C}\_\mu \leftarrow$ creating a dataset of centroids in C from epochs 1 to epoch $m - \mu$
    $[C\_F, C\_labels] \leftarrow kmeans(\mathbb{C}\_\mu, k)$ ►running clustering on $\mathbb{C}\_\mu$
  $labels \leftarrow$ Assigning each data point to its related centroid label
Return C_F, labels

4.4.4 Datasets

Clustering algorithms are developed either for a particular type of application or as a general solution that can be implemented in a variety of applications. In the first case, they are evaluated with one or more related datasets. For those that are proposed as a general solution, clustering methods should fulfil the following requirements to demonstrate their applicability as general methods:

1. The algorithm should be validated on various ranges of datasets.
2. The new algorithm should be compared with stable and widely used algorithms.
3. Datasets should be publicly available.
4. Experiments should avoid any type of bias.

To fulfil these criteria, the evaluation in the EFCA was designed to cover both multivariate and time series datasets, and the results were compared to the FCM and *K*-means, which are two of the most well-known and widely used algorithms. The datasets

utilised in this study are summarised in Table 4-4 and Table 4-5, and all are publicly available and accessible from (Dheeru, Dua and Karra Taniskidou, 2017). The only one that is not openly used is the Neutron and Gamma-ray dataset, which is a reduced dimensional version of the neutron and gamma-ray energy level. The experimental results of multivariate datasets and time series datasets are discussed in subsequent sections. In addition, the Gene Expression dataset was included in both categories of multivariate and time series datasets, because the UCI library classifies it as both types of datasets; it is thus included in both results for easier comparison with the results in each of these categories.

Table 4-4: Multivariate datasets

Dataset	**Instances**	**Attributes**
User Knowledge	403	5
Frog (Anuran Calls)	7,195	22
Iris	150	4
Seeds	210	7
Pulsars Star (HTRU2)	17,898	9
Gene Expression	801	20,531
Seizure	11,500	179
Wholesale	440	8

Table 4-5: Time series datasets

Dataset	Instances	Attributes
Gene Expression	801	20,531
Earthquakes	900	128
CBF	900	128
Electric Devices	322	512
Face All	1,690	131
Faces UCR	2,050	131
Inline Skate	550	1,882
Italy Power Demand	1,029	24
Phalanges Outline	1,800	80
Neutron and Gamma Ray	1,000	5

4.4.5 Obtained results and evaluation

In the following sections, the evaluation of the results obtained from multivariate and time series datasets is elaborated.

4.4.5.1 Multivariate datasets

Instant clustering approaches, regardless of the method they use for clustering, will cluster the whole dataset in a run and generate one set of cluster representatives. However, in evolving clustering, there are multiple epochs, and in each epoch, only a portion of the data will be clustered. In this method, a set of cluster representatives would exist for each epoch run; accordingly, the EFCA clusters the dataset using multiple sets of prototypes in various epochs. Furthermore, it introduces an “epoch cut” to develop better accuracy by cutting obscure data. In this study, there are five epochs for each dataset, so every epoch has 20% of data in it.

Table 4-6 demonstrates the obtained RI result - which is the indicator of accuracy - for the EFCA approach in comparison to the one obtained by *K*-means and the FCM method for eight multivariate datasets. The performance of the EFCA can be seen on the multivariate dataset with epoch cut = 0 in the fourth column, in which there is no epoch cut in datasets. The comparison of the EFCA with the FCM and *K*-means demonstrates that, except for User Knowledge dataset, which does not display better performance, the EFCA has almost better results for other datasets compared to the FCM and *K*-means. It has almost a 23% improvement compared to *K*-means on the Pulsars Star dataset by rising from 0.6423 to 0.8695, 12% on the Seizure dataset (from 0.5438 to 0.6635), 10% on the Frog dataset (from 0.7534 to 0.8487), and 4% on the Iris (0.8797 to 0.9146) and Seeds (from 0.8744 to 0.9149) datasets. Furthermore, it exhibits little improvement (about 2%) on the Wholesale and N&G datasets and almost the same improvement on the Gene Expression dataset. Likewise, the method demonstrated a 30% improvement in comparison with the FCM on the Pulsars Star dataset (from 0.5723 to 0.8695); 17% on the Seizure dataset (from 0.5002 to 0.6635); 10% on the Frog dataset (from 0.7437 to 0.8487); and 3.5–4.5% on the Iris (from 0.8797 to 0.9146), Seeds (from 0.8744 to 0.9149), and Wholesale (from 0.4898 to 0.5341) datasets, and it was almost the same for the Gene Expression and User Knowledge datasets.

Regarding the performance of the EFCA on multivariate datasets with one epoch cut, shown in the fifth column, the EFCA displays much better performance compared to the FCM approach, especially on the Frog, Iris, and Pulsars Star datasets, with an RI of 0.9757, 0.7460, and 1 respectively, which means a 23% improvement on the Frog dataset (from 0.7537 to 0.9757), 17% on the Pulsars Star dataset (from 05723 to 0.7460), 15% on the Iris dataset (from 0.8797 to 1), 12.5% on the Seizure dataset (from 0.5002 to 0.6251), and less than 5% on other datasets. Similarly, the EFCA outperforms *K*-means

on the Frog, Iris, Pulsars Star, and Seizure dataset at almost the same rate as the FCM in most of the datasets.

The performance of the EFCA on the multivariate dataset with two epoch cuts, shown in the sixth column, yielded an even better result for the EFCA, in which almost three RIs of 1 exist, thus demonstrating 100% accuracy in each cluster. This is an ideal result, considering that until now, achieving full accuracy was rare in clustering (although this seems to be too perfect to be real – this result is a consequence of a trade-off that the EFCA makes by eliminating obscure data in the last two epochs, which adds up to 40% of the data). The EFCA indicates more than a 22% improvement on the Frog dataset (from 0.7534 to 0.9789) and 12% on the Iris (from 0.8797 to 1) and Seeds (from 0.8744 to 1) datasets in comparison to the *K*-means method. Furthermore, in the rest of the datasets, it performs better or at least the same as the other two methods. On the other hand, the same is true in comparison with the FCM, in which results revealed even more accuracy in some cases.

The different epoch cuts are now compared on the quality of EFCA clustering when applied to multivariate data in the last three columns of Table 4-6. The positive effect of the increase in the number of epoch cuts is apparent, particularly in the Frog, Iris, and Seeds datasets with almost 100% accuracy, which connotes more than a 13% improvement. With the exception of the Pulsars Star dataset, increasing epoch cuts results in little fluctuation for the other datasets. It is apparent that the EFCA achieved significant improvements in clustering accuracy for some of these datasets, especially for those containing noisy data. This is a result of clustering the datasets using multiple sets of prototypes in various epochs. An “epoch cut” subsequently results in better accuracy by cutting obscure data in each cut. Furthermore, “post-pruning” automatically identifies and

eliminates obscure data. In Figure 4-13 to Figure 4-16, the results are also illustrated as bar charts in case the reader prefers a more visual comparison.

Table 4-6: Comparing EFCA clustering Rand index results with epoch cuts (EC) 0,1 and 2 with *K*-means and FCM on multivariate datasets

Multi-Variate Dataset Name	***K*-means RI**	**FCM RI**	**EFCA (EC = 0) RI**	**EFCA (EC = 1) RI**	**EFCA (EC = 2) RI**
User Knowledge	0.7292	0.7001	0.6959	0.7202	0.7109
Frog	0.7534	0.7437	0.8487	0.9757	0.9789
N&G	0.5878	0.5836	0.6115	0.6337	0.6337
Iris	0.8797	0.8797	0.9146	1	1
Seeds	0.8744	0.8744	0.9149	0.8997	1
Pulsars Star	0.6423	0.5723	0.8695	0.746	0.7132
Gene Expression	0.6548	0.6441	0.6523	0.6371	0.6455
Seizure	0.5438	0.5002	0.6635	0.6251	0.626
Wholesale	0.5168	0.4898	0.5341	0.5245	0.5425

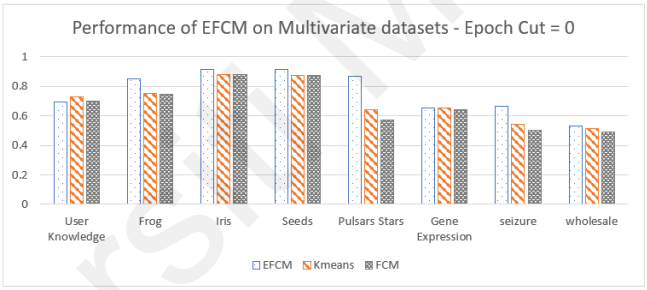

Figure 4-13: Performance of EFCA on multivariate dataset with no epoch cuts

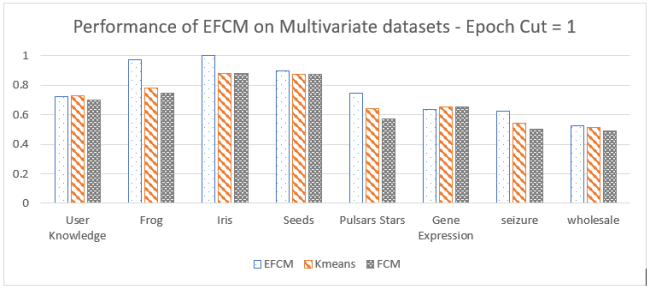

Figure 4-14: Performance of EFCA on multivariate dataset with one epoch cut

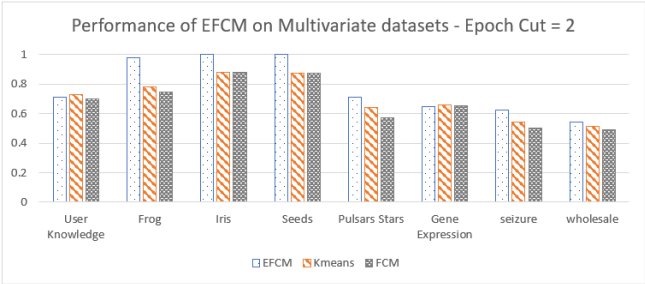

Figure 4-15: Performance of EFCA on multivariate dataset with two epoch cuts

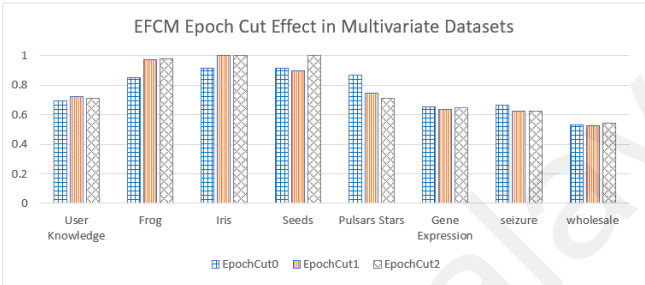

Figure 4-16: Comparison of different epoch cuts on the quality of EFCA clustering when applied to multivariate data

4.4.5.2 Time series datasets

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive, equally spaced points in time; thus, it is a sequence of discrete-time data.

The EFCA attempts to cluster those data with more explicit clusters at early epochs, and data that are clustered in final epochs can consequently be dismissed because of their obscurity. To achieve this aim, clustering is broken down into multiple epochs, and in each epoch, a membership matrix is used to cluster only those data points that have high similarity to prototypes; the rest of the data points are passed for further clustering in the next epoch. Table 4-7 illustrates the obtained RI results indicating the accuracy of clustering results by using the EFCA approach in comparison to *K*-means and the FCM method for 10 time series datasets.

The fourth column represents the performance of the EFCA on a time series dataset with no epoch cuts. The quality of clustering is demonstrated to be almost the same as that by *K*-means and the FCM in the Gene Expression, Inline Skate, Italy Power Demand, and Phalanges Outlines datasets. It is less than 5% better in the N&G (from 0.5836 to 0.6115) and CBF (from 0.701 to 0.7203) datasets, and 12% better in the Electric Devices dataset (from 0.6506 to 0. 7694) compared with the FCM approach. Furthermore, the EFCA has an excellent result on the Earthquake dataset (22–23% better) compared to *K*-means (from 0.5165 to 0.7386) and the FCM (from 0.5083 to 0.7386). It also has a much better performance in comparison with the FCM: an approximate 17% improvement in the Face All dataset (from 0.6559 to 0.8315) and a more than 25% improvement in the Face UCR dataset (from 0.579 to 0.831). However, *K*-means works slightly better (less than 6%) in these two datasets.

The performance of the EFCA on a time series dataset with one epoch cut, presented in the fifth column, indicates almost equal accuracy for the Gene Expression, Inline Skate, Italy Power Demand, and Phalanges Outlines datasets concerning these three methods. However, the EFCA not only performs somewhat better in the N&G and CBF datasets (from 0.701 to 0.7916), but also performs almost 11% better in the Electric Devices dataset (from 0.6506 to 0.7576), and it has much better performance in comparison with the FCM in the Face All (from 0.6559 to 0.7926) and Face UCR (from 0.579 to 0.7912) datasets: approximately 14% and 22% respectively.

The performance of the EFCA on a time series dataset with two epoch cuts, presented in the sixth column, reveals no difference between the accuracy of clustering based on RIs between these three methods on the Gene Expression, Inline Skate, and Italy Power Demand datasets. However, it performs over 31% better on the Earth Quake dataset (from 0.5083 to 0.8186), almost 15% better on the CBF dataset (from 0.701 to 0.8538), and 5%

better on the N&G dataset (0.5836 to 0.6337) with RIs than with the FCM. Such an improvement in accuracy is almost true in comparison with *K*-means as well. While in the Electric Devices, Face All, and Face UCR datasets, EFCA displays much better performance (from 0.6506, 0.6559, and 0.579 to 0.7235, 0.7245, and 0.7213 respectively) in comparison with FCM, which can be up to 15% in the last one, in the case of the Face All and Face UCR datasets, *K*-means works about 16% better than the EFCA (from 0.8875, and 0.8884 to 0.7245 and 0.7213).

A comparison of different epoch cuts on the quality of EFCA clustering when applied on time series data, as depicted in the last three columns, reveals that in the Gene Expression, Inline Skate, Italy Power Demand, and Phalanges Outlines datasets, there is little improvement (less than 5%) in performance accuracy with an increase in the number of epoch cuts, whereas for the N&G dataset, there is almost an 8% improvement. Furthermore, in the Earthquake (0.7386, 0.7455, 0.8186) and CBF (0.7203, 7916, 0.8538) datasets, there is much more improvement in accuracy (7-–14%). It is only in the Electric Devices (0.7694, 0.7576, 0.7235), Face All (0.8315, 0.7926, 0.7245), and Face UCR (0.831, 0.7212, 0.7213) datasets that the increase in the number of epoch cuts has a small backward effect (4.5% -10%). Overall, one can conclude that the epoch cut has resulted in better accuracy in more than 70% of the cases. In Figure 4-17 to Figure 4-20, the results are also demonstrated as bar charts for a visual comparison.

Correspondingly, another conclusion of these results confirms the much better performance of the EFCA in multivariate compared to time series datasets: a remarkable result of up to 100% accuracy can be obtained in some instances.

Table 4-7:Comparing EFCA clustering Rand index results with epoch cuts (EC) 0,1 and 2 with *K*-means and FCM on time series datasets

Time Series Dataset Name	*K*-means RI	FCM RI	EFCA (EC = 0) RI	EFCA (EC = 1) RI	EFCA (EC = 2) RI
Gene Expression	0.6543	0.6507	0.6527	0.6478	0.6365
Earthquakes	0.5165	0.5083	0.7386	0.7455	0.8186
CBF	0.7114	0.701	0.7203	0.7916	0.8538
Electric Devices	0.7572	0.6506	0.7694	0.7576	0.7235
Face All	0.8875	0.6559	0.8315	0.7926	0.7245
Faces UCR	0.8884	0.579	0.831	0.7912	0.7213
Inline Skate	0.7487	0.749	0.7371	0.7162	0.7453
Italy Power Demand	0.4999	0.5004	0.4996	0.5181	0.5
Phalanges Outline	0.5	0.5	0.5001	0.5057	0.5411
Neutron and Gamma Ray	0.5878	0.5836	0.6115	0.6337	0.6337

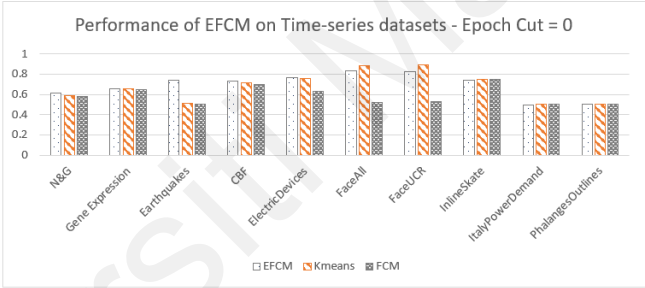

Figure 4-17: Performance of EFCA on time series dataset with no epoch cuts

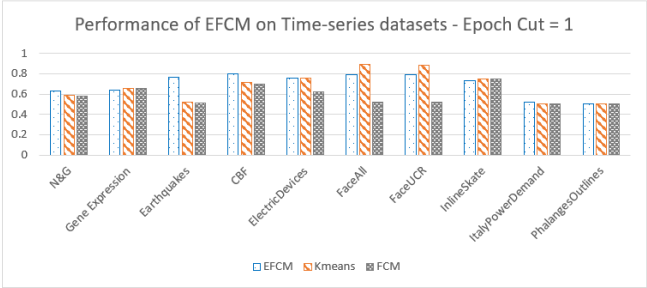

Figure 4-18: Performance of EFCA on time series dataset with one epoch cut

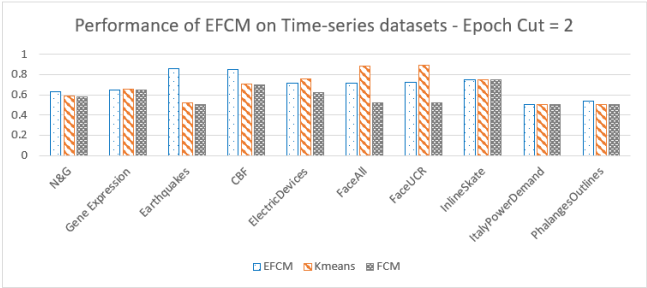

Figure 4-19: Performance of EFCA on Time series dataset with two epoch cuts

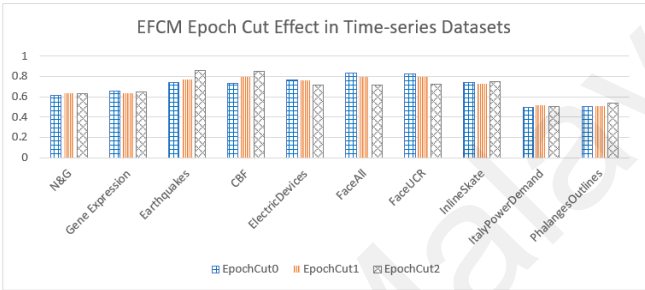

Figure 4-20: Comparison of different epoch cuts on the quality of EFCA clustering when applied on time series data

The visualised graphs in Figure 4-21 to Figure 4-25 help to make the results more crystalline. Based on the definition provided earlier, there are multiple epochs in the evolving clustering, and in each epoch, only a prescribed percentage of data will be picked based on the highest membership value to the cluster representatives that are available for each epoch run, as there are multiple sets of prototypes in multiple epochs. The remaining data points will be passed over to the next epoch for further clustering.

In Figure 4-21 to Figure 4-25, five epochs of the Iris dataset are considered with three clusters in each epoch as an example. The EFCA attempts to cluster all data with more explicit features in earlier epochs. In the first epoch, as illustrated in Figure 4-21, after clustering, 20% of the dataset with the highest membership value will be chosen. Then, the rest of the data that were clustered in that epoch will be passed down to the next epoch, and this process will be repeated in each epoch. The reason there are two clusters in the

first epoch is that these are the chosen data (20%) with the highest membership values in clusters in epoch one, and none of the data in this set belonged to the third cluster. The results of epochs 2, 3, 4, and 5 are presented in Figure 4-22 to Figure 4-25. In the fourth and fifth epochs, the clusters are not explicitly distinct from one another, and the borderlines of clusters are not clear. These figures visually demonstrate that data with explicit membership values are in the earlier epoch (1–3), and obscure data, which can be noisy data, are shifted to the last epochs (4–5). Considering this process, the exclusion of the last epoch can be expected to have a positive effect on the quality of the remaining data, and as a result, it can improve the quality and the accuracy of the next clusters.

Next, consider the Iris dataset with 150 data points. Each epoch has 20% of data, so 30 data points will be present in each epoch. After the exclusion of epochs 4 and 5 (60 data points), the remaining 90 data points will be clustered at the end in such a way that centroids will first be clustered in three groups, and the correspondent datapoints will then accompany them accordingly. As a result, after omission of ambiguous data, or noisy data, the remaining datapoints (3/5 of data) will be clustered during the final procedure, which will significantly increase the accuracy of clustering up to 100% in this dataset, as demonstrated in Table 4-6.

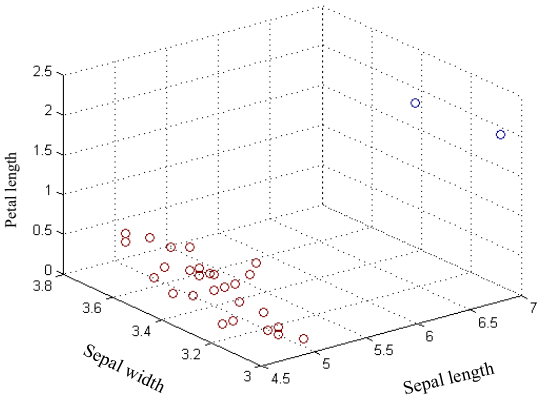

Figure 4-21: First epoch for Iris data

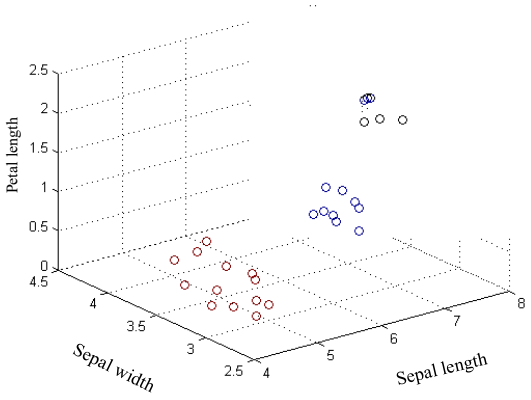

Figure 4-22: Second epoch for Iris data

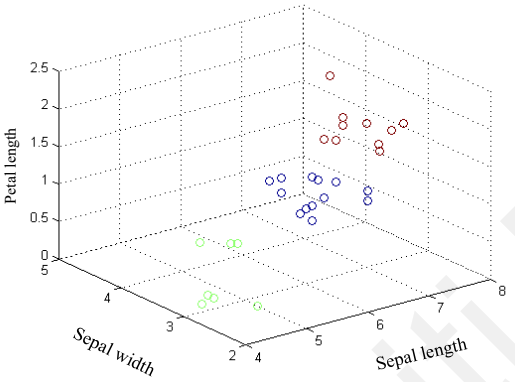

Figure 4-23: Third epoch for Iris data

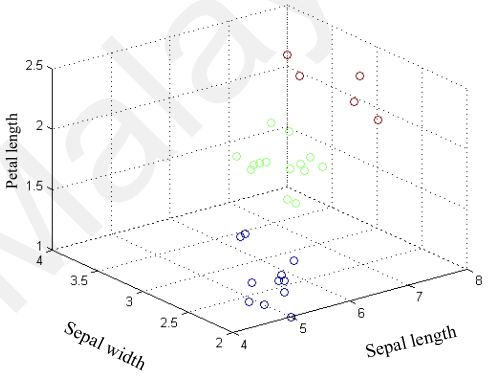

Figure 4-24: Fourth epoch for Iris data

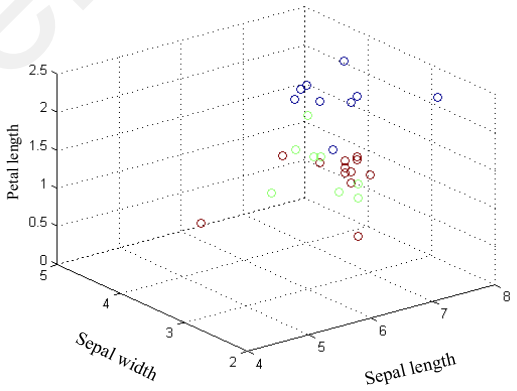

Figure 4-25: Fifth epoch for Iris data

The conclusion of these results confirms a much better performance of the EFCA in multivariate compared to time series datasets: a remarkable result of up to 100% accuracy can be obtained in some instances.

This characteristic of the EFCA would be advantageous in the big-data era nowadays. It can be especially useful as a pre-processing method for supervised methods, in which the quality and the accuracy of clustering is critical.

4.5 Significance of Findings

This study yielded many significant findings. They can be summarised as follows:

- The study, which was designed to determine which similarity or dissimilarity measure is most suitable for clustering continuous data overall, demonstrates that of all the similarity measures, the ED is one of the most suitable ones.
- Introducing an intelligent evolving clustering method demonstrates improvement in both multivariate and time series clustering accuracy. This method increased the accuracy of clustering on the total data (without any epoch cuts) up to 30% for multivariate datasets and up to 25% on time series data.
- The epoch cut notion, which enables heuristic data post-pruning to purify the data from noisy data, had a highly significant impact on multivariate data, resulting in an RI of 100%, which is perfect accuracy on some datasets. In time series datasets, epoch cuts resulted in an increase in accuracy of up to 31%.
- Apart from being a better clustering method in the unsupervised area, the introduced post-pruning method can also influence the area of supervised methods, by providing a smaller but, in some cases, significantly better training set.
- The EFCA also improved neutron and gamma-ray discrimination by 6% without epoch cuts and by 13% with them.

4.6 Summary of the Chapter

This chapter was divided into three phases, namely, similarity or dissimilarity measures, pre-processing, and the EFCA. Each section was dedicated to the detailed discussion of the procedure and evaluation of the proposed method. To ensure its generality, the EFCA was evaluated against a series of multivariate and time series datasets. Moreover, the observational results were evaluated to demonstrate how the EFCA outcomes were produced and how this method boosts the accuracy of the final clusters. The chapter then reported the results obtained by applying the EFCA to the datasets, and an extensive discussion of the results followed.

It has been confirmed that the EFCA outperforms other conventional clustering methods such as the FCM and *K*-means by testing on different datasets and demonstrating that accurate clusters are produced using multiple epochs and epoch cuts. The method proposed by the researcher can achieve the best results of clustering compared to several other popular methods.

CHAPTER 5: CONCLUSION

5.1 Introduction

This thesis launched with an investigation into the different types of clustering methods to investigate different well-known methods and represent their output and compare their quality of clustering. The proposed model was compared with existing traditional and well-known methods (the FCM and *K*-means). Several analyses of these methods were explored, and their capabilities were evaluated to satisfy the aims of this thesis. This chapter begins with a brief description of clustering and the method that is proposed, and then, in particular, it highlights the most important findings. Thereafter, it presents the main contributions of this thesis and elaborates on how the objectives were achieved. Finally, it concludes with a description of possible future work regarding the topics covered and the ways in which the proposed framework could be enhanced in the future.

5.2 Clustering Method

Clustering is an unsupervised machine learning method that is used both individually and as part of the pre-processing stage for supervised machine learning methods. The aim of clustering is to group similar data points into the same category, while grouping dissimilar data points into distinct clusters. Although clustering results can be used independently, in many instances, they are utilised as input for supervised learning methods. Given its unsupervised nature, clustering results in less accuracy compared to supervised learning. Therefore, most of the time, clustering results are not accurate enough and can result in unsuitable models. Considering this issue, having a smaller dataset with high clustering accuracy is preferable to having all datasets clustered with low accuracy.

Fuzzy clustering is different from hard clustering. In hard clusters, the item either is fit to a group or does not belong to it. In contrast, fuzzy clustering enables each item to pertain to several groups with some membership degrees from 0 to 1, depending on the distance between the item and cluster centres. In instant clustering approaches, regardless of the method they use for clustering, they will cluster the whole dataset in a run and generate one set of cluster representatives. This can also be described as dividing the dataset into various subgroups. This study presented a method for evolving fuzzy clustering that improves clustering accuracy by introducing a new perspective by using multiple sets of prototypes. To achieve this, clustering is broken down into multiple epochs instead of running the clustering on all data at once. In each epoch, a membership matrix is utilised to cluster only those data points that have high similarity to prototypes, and the rest of data points are passed for further clustering in the next epoch. Although much research has been carried out in the field of fuzzy clustering, this is the first time that membership values are used to generate multiple cluster prototypes. In the EFCA, prototypes are generated by a subset of data points in each epoch. Moreover, in each epoch, only a part of the dataset is clustered, unlike other fuzzy clustering methods that cluster all data at once. The EFCA employs one of the FCM features, which is the fuzzy membership matrix, to improve the clustering results. Then, instead of only one set of prototypes, the EFCA generates multiple sets of prototypes in multiple epochs and clusters the data gradually.

5.3 Summary of Results

This section summarises the major findings of the study by reviewing the achievements of the research.

5.3.1 Achievements of the study

The study proposed a novel, elaborative method with the overall goal of improving accuracy in clustering. The achievements of the study are summarised by answering the questions posed in Chapter 1.

The research questions answered in this thesis are listed below:

Q1. How does one develop a clustering approach that yields more accurate clustering results?

In some cases, for supervised learning, it is desirable to have a smaller subset of highly accurate labelled data instead of having all datasets labelled with low accuracy. The EFCA epoch cut provides this ability to concentrate on clustering the data points, which possibly results in more accuracy. In each epoch, the membership matrix is used to cluster only those data points that have high similarity to prototypes, and the rest of the data points are passed for further clustering in the next epoch. Furthermore, to automatically identify and disregard obscure examples, a heuristic approach, called "post-pruning", was introduced. The EFCA attempts to cluster more explicit data at early epochs, and later data clustered in final epochs can be disregarded because of their obscurity.

Q2. What is the influence of this method on the accuracy of continuous data (whether it is time series or multivariate)?

Clustering methods are either developed specifically for a certain type of application and tested on one or a set of related datasets, or they are developed as a general solution that can be deployed on a variety of datasets. In the latter, a

proposed clustering method should fulfil the following requirements to demonstrate its applicability as a general method: (1) the method must be validated on a different set of data, (2) the proposed method should be compared with stable and widely used methods, (3) datasets should be publicly available, and (4) experiments should avoid any type of bias. To fulfil these criteria, the experiments on the EFCA were designed to cover both multivariate and time series datasets, and the results were compared to the FCM and *K*-means, which are two of the most well-known and widely used methods.

Q 3. How can this method improve neutron and gamma-ray discrimination?

In nuclear physics, a critical procedure is the differentiation between neutron and gamma-ray pulses. Clustering will distinguish pulses based on their shapes from the scintillator detector. It consequently provides the flexibility to separate neutron particles from gamma rays abruptly, and nearly every pulse can thus be clustered as soon as it arrives. The EFCA clusters a dataset in multiple epochs using multiple sets of prototypes; it also introduces an "epoch cut" to improve accuracy by cutting obscure data. Then, a final round of clustering is conducted using the *K*-means method on cluster centres that are obtained in all epochs in order to attain the final result, which is the clustering of the primary dataset into two clusters.

5.3.2 Research objectives

The main objectives are as follows:

1. To develop a new method to clustering that is more accurate for neutron and gamma-ray pulses.

Discriminating neutrons from gamma rays is a vital task in nuclear physics and has been broadly used in different applications such as space research, mines, cultural heritage analysis, tomographical imaging, nuclear material control, international safeguarding, and national security. Traditional methods of discrimination have been analogue-based. However, ADCs have recently opened the door to the digital world and provided the possibility to analyse the pulses coming from scintillators by digital approaches. Digital pulses can be treated as time series, and clustering is a method to group time series based on their similarity in shape. Unlike the TOF approach, which is a widespread method for discrimination, the time series clustering approach does not need any data source other than pulses coming from the scintillator.

The pulses can consequently and rapidly be discriminated into neutron and gamma-ray, sometimes even immediately after they come from the scintillator. This study presented a method for evolving fuzzy clustering (namely, the EFCA) that improves clustering accuracy by introducing a new perspective in clustering through the use of multiple sets of prototypes. To cluster similar time series, a process of similarity matching must take place to calculate the similarity of whole time series. This process is known as whole time series clustering, in which the whole sequence of time series is studied when the distance is calculated. However, calculating similarity measures is not a simple task because of the noise and outliers. To overcome this problem, in the presented method, clustering is broken down into multiple epochs. Instead of running the clustering on all data at once, a membership matrix is used in each epoch to cluster only those data points that have high similarity to prototypes; the rest of the data points are passed for further

clustering in the next epoch. Afterward, the process of agglomeration takes place, and finally, since the problem has a crisp nature, which means a particle is either a neutron or a gamma-ray particle, *K*-means is used at the end of this procedure to discriminate neutron and gamma rays. The implementation of the newly proposed method was explained in Chapter 4, and it is the answer to the second question and the accomplishment of the first research objective.

2. To evaluate the capability of the suggested method for improving the accuracy of the clustering.

The comparison of the EFCA with the FCM and *K*-means demonstrates that the EFCA performs better than the other two methods in most datasets. It achieved significant improvements in clustering accuracy for some datasets; especially for those that have noisy data, an "epoch cut" resulted in better accuracy by cutting obscure data to the extent that for some datasets, the EFCA achieved 100% accuracy after epoch cut was applied. An epoch cut introduces a "post-pruning" method that can automatically identify and eliminate obscure examples. This was demonstrated in the experimental evaluation explicated in Chapter 4, and the second objective has thus been met.

3. To improve the performance of neutron and gamma-ray clustering (discrimination).

In this study, the EFCA clustering method was evaluated and compared to other well-known methods for its accuracy to indicate its influence on the discrimination task. The method exhibited a better result for the Neutron and Gamma-Ray dataset. The EFCA for this dataset generated a better result than *K*-

means and the FCM. Furthermore, the RI was used to evaluate the quality of the proposed method, and it was demonstrated that our proposed method has a higher quality compared to existing methods in most cases. This was proven in the experimental evaluation explained in Chapter 4, and it has addressed the third objective.

5.4 Significance of the Study

This study yielded a number of significant findings. They can be summarised as follows:

- To the best of our knowledge, for the first time, the literature review in this thesis study gathers and classifies the works that have been done in the area of fuzzy clustering.
- A thorough study was designed and carried out to determine which similarity/di-similarity is most suitable for clustering continuous data overall.
- The study introduces an intelligent evolving clustering method that demonstrates improvement in both multivariate and time series clustering accuracy.
- The study introduces epoch cuts, which enable heuristic data post-pruning to purify the data from noisy data.
- Apart from being a better clustering method in the unsupervised area, the introduced post-pruning method can also greatly influence the area of supervised methods by providing a smaller but, in some cases, significantly better training set.
- The study improved neutron and gamma-ray discrimination by 6% without epoch cuts and 13% with epoch cuts.

This research can change the way in which clustering algorithms are perceived and shift the focus from pre-processing methods to post-pruning approaches. There is still much room for improvement in both the epoch operation and the post-pruning components of the proposed method.

5.5 Limitation of Study

This study mainly focused on continuous data and did not cover categorical and binomial data, and the result might therefore not be extendable to these types of data. However, this might be a topic for future research works.

5.6 Future Works

The method that has been proposed in this study still has much of room for further studies. Some of these potential studies are listed below:

- Investigating the using of Type-II and intuitionistic fuzzy methods for evolution of the EFCA;
- Customising the EFCA for other applications such as streaming data and big data; and
- Utilising other distance measures for specific applications.

5.7 Conclusion and Further Research

The results of the experimental study suggest an unsupervised clustering method for clustering data that are affected by outliers or noise. This research implemented the EFCA on a set of multivariate and time series datasets to demonstrate the efficiency of the suggested method by experimental study. Experiments in this study indicate significant improvements in the clustering accuracy of some of the datasets, especially those containing noisy data. Furthermore, the results of the experimental study suggest

that the EFCA is a suitable unsupervised clustering method for clustering data affected by outliers or noise.

As mentioned above, this study mainly focused on introducing a new way of clustering based on the evolving concept by utilising a fuzzy membership matrix of the FCM that enables heuristic post-pruning to eliminate obscure or outlier data. However, a number of other clustering methods in the literature address the same problem. Therefore, comparing these methods with the proposed EFCA can be the subject of further research. Future works can also utilise the same concept, but with different fuzzy clustering methods for clustering in each epoch. Another research opportunity is to use a method other than *K*-means in the final stage to combine clustered data points in different epochs.

REFERENCES

Abd-Elaal, A. K., & Hefny, H. A. (2013). Forecasting of Egypt Wheat Imports Using Multivariate Fuzzy Time Series Model Based on Fuzzy Clustering. *IAENG International Journal of Computer Science*, *40*(4), 230–237. Retrieved from https://www.semanticscholar.org/paper/Forecasting-of-Egypt-Wheat-Imports-Using-Fuzzy-Time-Abd-Elaal-Hefny/fa2c02161f23d44f2db383ea8bf4ba2ee5f19543

Abdulla, S., & Al-Nassiri, A. (2015). kEFCM: kNN-Based Dynamic Evolving Fuzzy Clustering Method. *International Journal of Advanced Computer Science and Applications*, *6*(2), 5–13. https://doi.org/10.14569/IJACSA.2015.060202

Adams, J. M., & White, G. (1978). A versatile pulse shape discriminator for charged particle separation and its application to fast neutron time-of-flight spectroscopy. *Nuclear Instruments and Methods*, *156*(3), 459–476. https://doi.org/10.1016/0029-554X(78)90746-2

Aghabozorgi, S., Shirkhorshidi, A. S., Ying Wah, T., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering – A decade review. *Information Systems*, *53*, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Aghabozorgi, S., Ying Wah, T., Herawan, T., Jalab, H. A., Shaygan, M. A., & Jalali, A. (2014). A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique. *The Scientific World Journal*, *2014*, 562194. https://doi.org/10.1155/2014/562194

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient similarity search in sequence

databases. *Foundations of Data Organization and Algorithms*, *46*, 69–84.

Akkoyun, S. (2013). Time-of-flight discrimination between gamma-rays and neutrons by using artificial neural networks. *Annals of Nuclear Energy*, *55*, 297–301. https://doi.org/10.1016/j.anucene.2013.01.006

Alexander, T. K., & Goulding, F. S. (1961). An amplitude-insensitive system that distinguishes pulses of different shapes. *Nuclear Instruments and Methods*, *13*(C), 244–246. https://doi.org/10.1016/0029-554X(61)90198-7

Aljarah, I., Mafarja, M., Heidari, A. A., Faris, H., & Mirjalili, S. (2019). Multi-verse optimizer: Theory, literature review, and application in data clustering. In *Studies in Computational Intelligence* (Vol. 811, pp. 123–141). https://doi.org/10.1007/978-3-030-12127-3_8

Alonso, A. M., Berrendero, J. R., Hernández, A., & Justel, A. (2006). Time series clustering based on forecast densities. *Computational Statistics and Data Analysis*, *51*(2), 762–776. https://doi.org/10.1016/j.csda.2006.04.035

Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, *12*(4), 461–486. https://doi.org/10.1007/s10791-008-9066-8

Amiri, M., Přenosil, V., Cvachovec, F., Matěj, Z., & Mravec, F. (2014). Quick algorithms for real-time discrimination of neutrons and gamma rays. *Journal of Radioanalytical and Nuclear Chemistry*, *303*(1), 583–599. https://doi.org/10.1007/s10967-014-3406-5

Aparicio-Ruiz, P., Martín, E. B., Martín, J. G., & Achedad, P. C. (2019). Short-Term Forecasting in Office Consumption with Identification of Patterns by Clustering. In *Lecture Notes in Management and Industrial Engineering* (Ángel Orti, pp. 195–202). https://doi.org/10.1007/978-3-319-96005-0_24

Atanassov, K., & Gargov, G. (1989). Interval valued intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, *31*(3), 343–349. https://doi.org/10.1016/0165-0114(89)90205-4

Atanassov, K. T. (1986). Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, *20*(1), 87–96. https://doi.org/10.1016/S0165-0114(86)80034-3

Atanassov, Krassimir, & Georgiev, C. (1993). Intuitionistic fuzzy prolog. *Fuzzy Sets and Systems*, *53*(2), 121–128. https://doi.org/10.1016/0165-0114(93)90166-F

Babuška, R. (1998). Fuzzy Modeling for Control. In *Springer* (1st ed.). https://doi.org/10.1016/S1389-1723(02)80197-9

Back, H. O., Balata, M., Bellini, G., Benziger, J., Bonetti, S., Caccianiga, B., … Zuzel, G. (2008). Pulse-shape discrimination with the Counting Test Facility. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *584*(1), 98–113. https://doi.org/10.1016/J.NIMA.2007.09.036

Bagnall, A., & Janacek, G. (2005). Clustering time series with clipped data. *Machine Learning*, *58*(2–3), 151–178. https://doi.org/10.1007/s10994-005-5825-6

Banfield, J. D., & Raftery, A. E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, *49*(3), 803–821. https://doi.org/10.2307/2532201

Bao, D. (2008). A generalized model for financial time series representation and prediction. *Applied Intelligence*, *29*(1), 1–11. https://doi.org/10.1007/s10489-007-0063-1

Bao, D., & Yang, Z. (2008). Intelligent stock trading system by turning point confirming and probabilistic reasoning. *Expert Systems with Applications*, *34*(1), 620–627. https://doi.org/10.1016/j.eswa.2006.09.043

Barton, J. ., & Edgington, J. . (2000). Analysis of alpha-emitting isotopes in an inorganic scintillator. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *443*(2–3), 277–286. https://doi.org/10.1016/S0168-9002(99)01086-4

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In T. M. Kogan J., Nicholas C. (Ed.), *Grouping Multidimensional Data* (Vol. 2, pp. 25–71). https://doi.org/10.1007/3-540-28349-8_2

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2–3), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Bhaskar N. Patel, S. G. P. and D. K. I. L. (2012). Efficient Classification of Data Using Decision Tree. *Bonfring International Journal of Data Mining*, *2*(1), 13–16. https://doi.org/10.9756/BIJDM.1106

Birks, J. B., & Firk, F. W. K. (1965). The Theory and Practice of Scintillation Counting. *Physics Today*, *18*(8). https://doi.org/10.1063/1.3047620

Bode, G., Schreiber, T., Baranski, M., & Müller, D. (2019). A time series clustering approach for Building Automation and Control Systems. *Applied Energy*, *238*, 1337–1345. https://doi.org/10.1016/j.apenergy.2019.01.196

Bonner, R. E. (1964). On Some Clustering Techniques. *IBM Journal of Research and Development*, *8*(1), 22–32. https://doi.org/10.1147/rd.81.0022

Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. *In Proceedings of the Eighth SIAM International Conference on Data Mining*, *1*, 243–254. https://doi.org/10.1137/1.9781611972788.22

Brooks, F. D. (1959). A scintillation counter with neutron and gamma-ray discriminators. *Nuclear Instruments and Methods*, *4*(3), 151–163. https://doi.org/10.1016/0029-554X(59)90067-9

Brooks, F. D. (1979). Development of organic scintillators. *Nuclear Instruments and Methods*, *162*(1–3), 477–505. https://doi.org/10.1016/0029-554X(79)90729-8

Budakovsky, S. V., Galunov, N. Z., Grinyov, B. V., Karavaeva, N. L., Kyung Kim, J., Kim, Y.-K., … Tarasenko, O. A. (2007). Stilbene crystalline powder in polymer base as a new fast neutron detector. *Radiation Measurements*, *42*(4–5), 565–568. https://doi.org/10.1016/J.RADMEAS.2007.02.058

Bustince, H., & Burillo, P. (1995). Correlation of interval-valued intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, *74*(2), 237–244. https://doi.org/10.1016/0165-0114(94)00343-6

Bustince, H., & Burillo, P. (1996). Vague sets are intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, *79*(3), 403–405. https://doi.org/10.1016/0165-0114(95)00154-9

Cai, W., Chen, S., & Zhang, D. (2007). Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognition*, *40*(3), 825–838. https://doi.org/10.1016/j.patcog.2006.07.011

Castillo, O., & Atanassov, K. (2019). Comments on Fuzzy Sets, Interval Type-2 Fuzzy Sets, General Type-2 Fuzzy Sets and Intuitionistic Fuzzy Sets. In C. O. Melliani S. (Ed.), *Recent Advances in Intuitionistic Fuzzy Logic Systems* (pp. 35–43). https://doi.org/10.1007/978-3-030-02155-9_3

Cha, & Sung-Hyuk. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *1*(4), 300–307. https://doi.org/10.1.1.154.8446

Chaira, T. (2010). A novel intuitionistic fuzzy C means clustering algorithm and its application to medical images. *Applied Soft Computing*, *11*(2), 1711–1717. https://doi.org/10.1016/j.asoc.2010.05.005

Chen, C. H., Pau, L. F., Wang, P. S. P., & Dubes, R. C. (1993). Cluster analysis and related issues. In *Handbook of Pattern Recognition and Computer Vision* (pp. 3–32). https://doi.org/10.1142/9789814343138_0001

Chen, L., Guo, G., & Wang, S. (2012). Nearest Neighbor Classification by Partially Fuzzy Clustering. *2012 26th International Conference on Advanced Information Networking and Applications Workshops*, 789–794.

https://doi.org/10.1109/WAINA.2012.23

Chen, Q., Sun, J., Palade, V., Shi, X., & Liu, L. (2019). Hierarchical Clustering Based Band Selection Algorithm for Hyperspectral Face Recognition. *IEEE Access*, *7*, 24333–24342. https://doi.org/10.1109/ACCESS.2019.2897213

Chintalapudi, K. K., & Kam, M. (1998). A noise-resistant fuzzy c means algorithm for clustering. *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36228)*, *2*, 1458–1463. https://doi.org/10.1109/FUZZY.1998.686334

Choudhry, M. S., & Kapoor, R. (2016). Performance Analysis of Fuzzy C-Means Clustering Methods for MRI Image Segmentation. *Procedia Computer Science*, *89*, 749–758. https://doi.org/10.1016/j.procs.2016.06.052

Coppi, R., D'Urso, P., & Giordani, P. (2006). Fuzzy C-medoids clustering models for time-varying data. In *Modern Information Processing* (pp. 195–206). https://doi.org/10.1016/B978-044452075-3/50017-0

Cumming, G. (2011). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis* (1, Ed.). Retrieved from https://books.google.com/books?hl=en&lr=&id=1W6laNc7Xt8C&oi=fnd&pg=PR1&dq=Understanding+The+New+Statistics:+Effect+Sizes,+Confidence+Intervals,+and+Meta-Analysis&ots=PuHRVGc55O&sig=cEg6l3tSxFHlTI5dvubr1j7yMpI

D'Mellow, B., Aspinall, M. D., Mackin, R. O., Joyce, M. J., & Peyton, A. J. (2007). Digital discrimination of neutrons and gamma-rays in liquid scintillators using pulse gradient analysis. *Nuclear Instruments and Methods in Physics Research, Section A:*

Accelerators, Spectrometers, Detectors and Associated Equipment, *578*(1), 191–197. https://doi.org/10.1016/j.nima.2007.04.174

D'Urso, P. (2005). Fuzzy clustering for data time arrays with inlier and outlier time trajectories. *IEEE Transactions on Fuzzy Systems*, *13*(5), 583–604. https://doi.org/10.1109/TFUZZ.2005.856565

D'Urso, P., De Giovanni, L., & Massari, R. (2015). Time series clustering by a robust autoregressive metric with application to air pollution. *Chemometrics and Intelligent Laboratory Systems*, *141*, 107–124. https://doi.org/10.1016/j.chemolab.2014.11.003

D'Urso, P., & Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*, *160*(24), 3565–3589. https://doi.org/10.1016/j.fss.2009.04.013

Dave, R. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, *12*(11), 657–664. https://doi.org/10.1016/0167-8655(91)90002-4

Dave, R. (1993). Robust fuzzy clustering algorithms. *[Proceedings 1993] Second IEEE International Conference on Fuzzy Systems*, *2*, 1281–1286. https://doi.org/10.1109/FUZZY.1993.327577

Daxin Jiang, Chun Tang, & Aidong Zhang. (2004). Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, *16*(11), 1370–1386. https://doi.org/10.1109/TKDE.2004.68

De, S. K., Biswas, R., & Roy, A. R. (2000). Some operations on intuitionistic fuzzy sets.

Fuzzy Sets and Systems, *114*(3), 477–484. https://doi.org/10.1016/S0165-0114(98)00191-2

Deschrijver, G., Cornelis, C., & Kerre, E. E. (2004). On the Representation of Intuitionistic Fuzzy t-Norms and t-Conorms. *IEEE Transactions on Fuzzy Systems*, *12*(1), 45–61. https://doi.org/10.1109/TFUZZ.2003.822678

Deschrijver, G., & Kerre, E. E. (2003). On the relationship between some extensions of fuzzy set theory. *Fuzzy Sets and Systems*, *133*(2), 227–235. https://doi.org/10.1016/S0165-0114(02)00127-6

Deschrijver, G., & Kerre, E. E. (2007). On the position of intuitionistic fuzzy set theory in the framework of theories modelling imprecision. *Information Sciences*, *177*(8), 1860–1866. Retrieved from https://www.sciencedirect.com/science/article/pii/S0020025506003410

Deshmukh, R., & Hwang, I. (2019). Anomaly Detection Using Temporal Logic Based Learning for Terminal Airspace Operations. *AIAA Scitech 2019 Forum*, 0682. https://doi.org/10.2514/6.2019-0682

Dheeru, Dua and Karra Taniskidou, E. (2017). UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. Querying and mining of time series data. *Proceedings of the VLDB Endowment*, *1*(2), 1542–1552. https://doi.org/10.14778/1454159.1454226

Dinh Nguyen, D., Ngo, L. T., & Pham, L. T. (2013). GMKIT2-FCM: A Genetic-based

improved Multiple Kernel Interval Type-2 FUzzy C-means clustering. *2013 IEEE International Conference on Cybernetics (CYBCO)*, 104–109. https://doi.org/10.1109/CYBConf.2013.6617457

Doucet, E., Brown, T., Chowdhury, P., Lister, C. J., Morse, C., Bender, P. C., & Rogers, A. M. (2018). Machine learning n/γ discrimination in CLYC scintillators. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *954*, 161201. https://doi.org/10.1016/J.NIMA.2018.09.036

Dunham, M. H. (2003). *Data Mining Introductor and Advanced Topics*. Upper Saddle River, New Jersey: Prentice Hall.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, *3*(3), 32–57. https://doi.org/10.1080/01969727308546046

Duru, O., & Bulut, E. (2014). A non-linear clustering method for fuzzy time series: Histogram damping partition under the optimized cluster paradox. *Applied Soft Computing*, *24*, 742–748. https://doi.org/10.1016/j.asoc.2014.08.038

E Gatti, F. M. (1962). A new linear method of discrimination between elementary particles in scintillation counters. *Nuclear Electronics II. Proceedings of the Conference on Nuclear Electronics. V. II.* Retrieved from https://inis.iaea.org/search/search.aspx?orig_q=RN:43116654

Everitt, B., Landau, S., & Leese, M. (2011). *Cluster analysis.* (5th ed.). Retrieved from https://www.wiley.com/en-us/Cluster+Analysis%2C+5th+Edition-p-

9780470749913

Faloutsos, C., Ranganathan, M., & Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases. *ACM SIGMOD Record*, *23*(2), 419–429.

Fisher, R. A. (1992). Statistical Methods for Research Workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics* (pp. 66–70). https://doi.org/10.1007/978-1-4612-4380-9_6

Fowlkes, E., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*(383), 553–569.

Fu, L., & Medico, E. (2007). FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics*, *8*(1), 3. https://doi.org/10.1186/1471-2105-8-3

Gan, G., Ma, C., & Wu, J. (2007). Data Clustering theory, Algorithms, and Applications. In *ASASIAM Series on Statistics and Applied*. Society for Industrial and Applied Mathematics.

Gath, I., & Geva, A. B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 773–780. https://doi.org/10.1109/34.192473

Gau, W. L., & Buehrer, D. J. (1993). Vague Sets. *IEEE Transactions on Systems, Man and Cybernetics*, *23*(2), 610–614. https://doi.org/10.1109/21.229476

Gerstenkorn, T., & Mańko, J. (1991). Correlation of intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, *44*(1), 39–43. https://doi.org/10.1016/0165-0114(91)90031-K

Gionis, A., Mannila, H., & Tsaparas, P. (2005). Clustering aggregation. *ACM Trans. Knowl. Discov. Data*, *1*(1), Article 4. https://doi.org/10.1109/ICDE.2005.34

Goguen, J. . (1967). L-fuzzy sets. *Journal of Mathematical Analysis and Applications*, *18*(1), 145–174. https://doi.org/10.1016/0022-247X(67)90189-8

Gordon, A. D. (1999). *Classification* (2nd ed.). Retrieved from https://www.crcpress.com/Classification/Gordon/p/book/9781584880134

Gosain, A., & Dahiya, S. (2016). Performance Analysis of Various Fuzzy Clustering Algorithms: A Review. *Procedia Computer Science*, *79*(6), 100–111. https://doi.org/10.1016/j.procs.2016.03.014

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857–871. Retrieved from http://www.jstor.org/stable/2528823

Guan, C., Yuen, K. K. F., & Coenen, F. (2019). Particle swarm Optimized Density-based Clustering and Classification: Supervised and unsupervised learning approaches. *Swarm and Evolutionary Computation*, *44*, 876–896. https://doi.org/10.1016/J.SWEVO.2018.09.008

Gullo, F., Ponti, G., Tagarelli, A., Tradigo, G., & Veltri, P. (2012). A time series approach for clustering mass spectrometry data. *Journal of Computational Science*, *3*(5), 344–355. https://doi.org/10.1016/j.jocs.2011.06.008

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. In R. Meersman, Z. Tari, & D. C. Schmidt (Eds.), *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003.*

Lecture Notes in Computer Science (pp. 986–996). https://doi.org/10.1007/978-3-540-39964-3_62

Gustafson, D., & Kessel, W. (2008). Fuzzy clustering with a fuzzy covariance matrix. *1978 IEEE Conference on Decision and Control Including the 17th Symposium on Adaptive Processes*, 761–766. https://doi.org/10.1109/cdc.1978.268028

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information*, *17*(2), 107–145.

Hamel, M., Sibczynski, P., Blanc, P., Iwanowska, J., Carrel, F., Syntfeld-Każuch, A., & Normand, S. (2014). A fluorocarbon plastic scintillator for neutron detection: Proof of concept. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *768*, 26–31. https://doi.org/10.1016/J.NIMA.2014.09.029

Han, J, Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*. Morgan Kaufmann.

Han, Jiawei., Kamber, M., & Pei, J. (2011). *Data mining : concepts and techniques*. Retrieved from https://books.google.com.my/books?hl=en&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=Data+Mining:+Concepts+and+Techniques&ots=tzIuZSryVW&sig=HjlifmBy6Iy-icbrhH3m9u5DsYI&redir_esc=y#v=onepage&q=Data Mining%3A Concepts and Techniques&f=false

Hand, D., Mannila, H., & Smyth, P. (2001). Principles of data mining(adaptive computation and machine learning). In *Drug safety*.

Herrera, F., Martínez, L., & Sánchez, P. J. (2005). Managing non-homogeneous information in group decision making. *European Journal of Operational Research*, *166*(1), 115–132. https://doi.org/10.1016/j.ejor.2003.11.031

Hong, D. H., & Choi, C. H. (2000). Multicriteria fuzzy decision-making problems based on vague set theory. *Fuzzy Sets and Systems*, *114*(1), 103–113. https://doi.org/10.1016/S0165-0114(98)00271-1

Hong, D. H., & Hwang, S. Y. (1995). Correlation of intuitionistic fuzzy sets in probability spaces. *Fuzzy Sets and Systems*, *75*(1), 77–81. https://doi.org/10.1016/0165-0114(94)00330-A

Höppner, F., & Klawonn, F. (2009). Compensation of translational displacement in time series clustering using cross correlation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5772 LCNS*, 71–82. https://doi.org/10.1007/978-3-642-03915-7_7

Hruschka, E. R., Campello, R. J. G. B., Freitas, A. A., & de Carvalho, A. C. P. L. F. (2009a). A Survey of Evolutionary Algorithms for Clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *39*(2), 133–155. https://doi.org/10.1109/TSMCC.2008.2007252

Hruschka, E. R., Campello, R. J. G. B., Freitas, A. A., & de Carvalho, A. C. P. L. F. (2009b). A Survey of Evolutionary Algorithms for Clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *39*(2), 133–155. https://doi.org/10.1109/TSMCC.2008.2007252

Hu, J., Pan, L., Yang, Y., & Chen, H. (2019). A group medical diagnosis model based on

intuitionistic fuzzy soft sets. *Applied Soft Computing*, *77*, 453–466. https://doi.org/10.1016/J.ASOC.2019.01.041

Huang, H. C., Chuang, Y. Y., & Chen, C. S. (2012). Multiple kernel fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, *20*(1), 120–134. https://doi.org/10.1109/TFUZZ.2011.2170175

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

HUNG, W.-L. (2003). Using statistical viewpoint in developing correlation of intuitionistic fuzzy sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *9*(4), 509–516. https://doi.org/10.1142/s0218488501000910

Hung, W. L., & Wu, J. W. (2002). Correlation of intuitionistic fuzzy sets by centroid method. *Information Sciences*, *144*(1–4), 219–225. https://doi.org/10.1016/S0020-0255(02)00181-0

Hung, W. L., & Yang, M. S. (2007). Similarity measures of intuitionistic fuzzy sets based on Lp metric. *International Journal of Approximate Reasoning*, *46*(1), 120–136. https://doi.org/10.1016/j.ijar.2006.10.002

Hwang, C., & Rhee, F. C.-H. (2007). Uncertain Fuzzy Clustering: Interval Type-2 Fuzzy Approach to c-Means. *IEEE Transactions on Fuzzy Systems*, *15*(1), 107–120. https://doi.org/10.1109/TFUZZ.2006.889763

Izakian, H., & Pedrycz, W. (2014a). Agreement-based fuzzy C-means for clustering data

with blocks of features. *Neurocomputing*, *127*, 266–280. https://doi.org/10.1016/j.neucom.2013.08.006

Izakian, H., & Pedrycz, W. (2014b). Anomaly Detection and Characterization in Spatial Time Series Data: A Cluster-Centric Approach. *IEEE Transactions on Fuzzy Systems*, *22*(6), 1612–1624. https://doi.org/10.1109/TFUZZ.2014.2302456

Izakian, H., Pedrycz, W., & Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, *39*, 235–244. https://doi.org/10.1016/j.engappai.2014.12.015

Izakian, H., Pedrycz, W., Jamal, I., Hesam, I., Pedrycz, W., Jamal, I., … Jamal, I. (2013). Clustering Spatiotemporal Data: An Augmented Fuzzy C-Means. *IEEE Transactions on Fuzzy Systems*, *21*(5), 855–868. https://doi.org/10.1109/TFUZZ.2012.2233479

Jain, A., & Dubes, R. (1988). Algorithms for clustering data. In *Technometrics* (Vol. 32). https://doi.org/10.2307/1268876

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Jain, A. K., Murty, M. N., Fynn, P. J., A.K., J., M.N., M., & P.J., F. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323.

János Abonyi, B. F. (2007). *Cluster Analysis for Data Mining and System Identification*. Springer.

Jantzen, J., Norup, J., Dounias, G., & Bjerregaard, B. (2005). Pap-smear Benchmark Data

For Pattern Classification. *Proc. NiSIS 2005, Albufeira, Portugal*, 1–9. Retrieved from http://fuzzy.iau.dtu.dk/download/smear2005

Jeremy, M., & Diansheng, G. (2009). Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems*, *33*(6), 403–408. https://doi.org/10.1016/J.COMPENVURBSYS.2009.11.001

Ji, M., Xie, F., & Ping, Y. (2013). A Dynamic Fuzzy Cluster Algorithm for Time Series. *Abstract and Applied Analysis*, *2013*, 1–7. https://doi.org/10.1155/2013/183410

Jin, D., & Bai, X. (2019). Distribution Information Based Intuitionistic Fuzzy Clustering for Infrared Ship Segmentation. *IEEE Transactions on Fuzzy Systems*, 1–1. https://doi.org/10.1109/TFUZZ.2019.2917809

Joyce, M. J., Aspinall, M. D., Cave, F. D., Georgopoulos, K., & Jarrah, Z. (2010). The design, build and test of a digital analyzer for mixed radiation fields. *IEEE Transactions on Nuclear Science*, *57*(5), 2625–2630. https://doi.org/10.1109/TNS.2010.2044245

Kadam, S., & Appl, D. T.-I. J. C. T. (2012). High dimensional data mining in time series by reducing dimensionality and numerosity. *International Journal of Computer Technology & Applications*, *3*(3), 903–908. Retrieved from https://pdfs.semanticscholar.org/0781/c5585debc6cdbf6c92340f1bcfa66fafa45f.pdf

Kannan, S. R., Ramathilagam, S., & Chung, P. C. (2012). Effective fuzzy c-means clustering algorithms for data clustering problems. *Expert Systems with Applications*, *39*(7), 6292–6300. https://doi.org/10.1016/j.eswa.2011.11.063

Kant, N., & Mahajan, M. (2019). Time-Series Outlier Detection Using Enhanced K-Means in Combination with PSO Algorithm. In *Lecture Notes in Electrical Engineering* (Vol. 478, pp. 363–373). https://doi.org/10.1007/978-981-13-1642-5_33

Kapitanov, A., Kapitanova, I., Troyanovskiy, V., Ilyushechkin, V., & Dorogova, E. (2019). Clustering of Word Contexts as a Method of Eliminating Polysemy of Words. *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 1861–1864. https://doi.org/10.1109/EIConRus.2019.8656851

Kasabov, N. K., & Qun Song. (2002). DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction. *IEEE Transactions on Fuzzy Systems*, *10*(2), 144–154. https://doi.org/10.1109/91.995117

Kaschuck, Y., & Esposito, B. (2005). Neutron/γ-ray digital pulse shape discrimination with organic scintillators. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *551*(2–3), 420–428. https://doi.org/10.1016/J.NIMA.2005.05.071

Kaur, P., & Gosain, A. (2010). Density-oriented approach to identify outliers and get noiseless clusters in Fuzzy C-Means. *International Conference on Fuzzy Systems*, 1–8. https://doi.org/10.1109/FUZZY.2010.5584592

Kaur, P., Soni, A. K., & Gosain, A. (2011). Robust Intuitionistic Fuzzy C-means clustering for linearly and nonlinearly separable data. *2011 International Conference on Image Information Processing*, *11*(3), 1–6.

https://doi.org/10.1109/ICIIP.2011.6108908

Kaur, P., Soni, A. K., & Gosain, A. (2013). Robust kernelized approach to clustering by incorporating new distance measure. *Engineering Applications of Artificial Intelligence*, *26*(2), 833–847. https://doi.org/10.1016/J.ENGAPPAI.2012.07.002

Kaur, P., Soni, A. K., Gosain, A., Soni, D. A. K., & Anjana Gosain, D. R. (2012). Novel intuitionistic fuzzy c-means clustering for linearly and nonlinearly separable data. *WSEAS Transactions on Computers*, *11*(3), 65–76. https://doi.org/10.1109/ICIIP.2011.6108908

Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, *7*(4), 349–371.

Keogh, E., Pazzani, M., Chakrabarti, K., & Mehrotra, S. (2000). A simple dimensionality reduction technique for fast similarity search in large time series databases. *Knowledge and Information Systems*, *1805*(1), 122–133.

Keogh, E., & Ratanamahatana, C. (2004). Exact indexing of dynamic time warping. *Knowledge and Information Systems*, *7*(3), 358–386. https://doi.org/10.1007/s10115-004-0154-9

Khalifi, H., Cherif, W., Qadi, A. El, & Ghanou, Y. (2019). Query expansion based on clustering and personalized information retrieval. *Progress in Artificial Intelligence*, *8*(2), 241–251. https://doi.org/10.1007/s13748-019-00178-y

Killick, R., Eckley, I. A., Ewans, K., & Jonathan, P. (2010). Detection of changes in

variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, *37*(13), 1120–1126. https://doi.org/10.1016/j.oceaneng.2010.04.009

Kim, H., Kim, H. K., Kim, M., Park, J., Cho, S., Im, K. Bin, & Ryu, C. R. (2019). Representation learning for unsupervised heterogeneous multivariate time series segmentation and its application. *Computers and Industrial Engineering*, *130*, 272–281. https://doi.org/10.1016/j.cie.2019.02.029

Knoll, G. F. (2010). *Radiation detection and measurement* (4th, Ed.). Retrieved from https://books.google.com.my/books?hl=en&lr=&id=4vTJ7UDel5IC&oi=fnd&pg=PA1&dq=G.F.+Knoll,+Radiation+Detection+and+Measurement,+3rd+ed.,+Wiley,+New+York,+2001.&ots=VwDTGTxJey&sig=0jYwVHvOt9WCgZ1wf-zH36xjgZA&redir_esc=y#v=onepage&q&f=false

Kolatch, E. (2001). Clustering algorithms for spatial databases: a survey. *PDF Is Available on the Web*, 1–22. Retrieved from https://pdfs.semanticscholar.org/4db2/a23114110c0d9c586c02737508f3ec71ee26.pdf

Kornilov, N. V., Fabry, I., Oberstedt, S., & Hambsch, F.-J. (2009). Total characterization of neutron detectors with a 252Cf source and a new light output determination. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *599*(2–3), 226–233. https://doi.org/10.1016/J.NIMA.2008.10.032

Kornilov, N. V., Khriatchkov, V. A., Dunaev, M., Kagalenko, A. B., Semenova, N. N., Demenkov, V. G., & Plompen, A. J. M. (2003). Neutron spectroscopy with fast

waveform digitizer. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *497*(2–3), 467–478. https://doi.org/10.1016/S0168-9002(02)01790-4

Kousar, A., Mittal, N., & Singh, P. (2020). An Improved Hierarchical Clustering Method for Mobile Wireless Sensor Network Using Type-2 Fuzzy Logic. *Lecture Notes in Electrical Engineering*, *605*, 128–140. https://doi.org/10.1007/978-3-030-30577-2_11

Krishnapuram, R., & Keller, J. M. (1996). The possibilistic C-means algorithm: insights and recommendations. *IEEE Transactions on Fuzzy Systems*, *4*(3), 385–393. https://doi.org/10.1109/91.531779

Krishnapuram, Raghu, & Keller, J. M. (1993a). A Possibilistic Approach to Clustering. *IEEE Transactions on Fuzzy Systems*, *1*(2), 98–110. https://doi.org/10.1109/91.227387

Krishnapuram, Raghu, & Keller, J. M. (1993b). Possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, *1*(2), 98–110. https://doi.org/10.1109/91.227387

Kumar, D., Agrawal, R. K., & Singh Kirar, J. (2019). Intuitionistic Fuzzy Clustering Method with Spatial Information for MRI Image Segmentation. *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, *2019-June*, 1–7. https://doi.org/10.1109/FUZZ-IEEE.2019.8858865

Kumar, D., Verma, H., Mehra, A., & Agrawal, R. K. (2019). A modified intuitionistic fuzzy c-means clustering approach to segment human brain MRI image. *Multimedia Tools and Applications*, *78*(10), 12663–12687. https://doi.org/10.1007/s11042-018-

5954-0

Legendre, P., & Legendre, L. (2012). *Numerical ecology*. Elsevier.

Li, C. S., Wang, Y., & Yang, H. (2010). Combining fuzzy partitions using fuzzy majority vote and KNN. *Journal of Computers*, *5*(5), 791–798. https://doi.org/10.4304/jcp.5.5.791-798

Li, D., Gu, H., & Zhang, L. (2010). A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data. *Expert Systems with Applications*, *37*(10), 6942–6947. https://doi.org/10.1016/J.ESWA.2010.03.028

Li, J., & Lewis, H. W. (2016). Fuzzy Clustering Algorithms — Review of the Applications. *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, 282–288. https://doi.org/10.1109/SmartCloud.2016.14

Li, Y., Olson, D. L., & Qin, Z. (2007). Similarity measures between intuitionistic fuzzy (vague) sets: A comparative analysis. *Pattern Recognition Letters*, *28*(2), 278–285. https://doi.org/10.1016/j.patrec.2006.07.009

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern Recognition*, *38*(11), 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025

Liao, T. W., & Warren Liao, T. (2005). Clustering of time series data - A survey. *Pattern Recognition*, *38*(11), 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025

Lin, Duan, & Tian. (2020). Interval Intuitionistic Fuzzy Clustering Algorithm Based on Symmetric Information Entropy. *Symmetry*, *12*(1), 79. https://doi.org/10.3390/sym12010079

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '03*, 2–11. https://doi.org/10.1145/882085.882086

Lin, J., Vlachos, M., Keogh, E., & Gunopulos, D. (2004). Iterative incremental clustering of time series. *Advances in Database Technology-EDBT 2004*, 521–522.

Lin, Jessica, Keogh, E., Lonardi, S., Lankford, J. P., & Nystrom, D. M. (2004). Visually mining and monitoring massive time series. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04*, 460. https://doi.org/10.1145/1014052.1014104

Lin, Jessica, Keogh, E., Lonardi, S., & Patel, P. (2002). *Finding Motifs in Time Series*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.6629&rep=rep1&type=pdf

Lin, Jessica, Keogh, E., & Truppel, W. (2003). Clustering of streaming time series is meaningless. *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '03*, 56–65. https://doi.org/10.1145/882082.882096

Lin, Jessica, Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, *15*(2), 107–144. https://doi.org/10.1007/s10618-007-0064-z

Lin, K. P. (2014). A Novel Evolutionary Kernel Intuitionistic Fuzzy C-means Clustering

Algorithm. *IEEE Transactions on Fuzzy Systems*, *22*(5), 1074–1087. https://doi.org/10.1109/TFUZZ.2013.2280141

Lin, S., Song, M., & Zhang, L. (2008). Comparison of cluster representations from partial second-to full fourth-order cross moments for data stream clustering. *Eighth IEEE International Conference on Data Mining, 2008. ICDM '08*, 560–569.

Linda, O., & Manic, M. (2012). General Type-2 Fuzzy C-Means Algorithm for Uncertain Fuzzy Clustering. *Ieee Transactions on Fuzzy Systems*, *20*(5), 883–897. https://doi.org/Doi 10.1109/Tfuzz.2012.2187453

Liu, G., Aspinall, M. D., Ma, X., & Joyce, M. J. (2009). An investigation of the digital discrimination of neutrons and γ rays with organic scintillation detectors using an artificial neural network. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *607*(3), 620–628. https://doi.org/10.1016/j.nima.2009.06.027

Liu, Guofu, Joyce, M. J., Ma, X., & Aspinall, M. D. (2010). A Digital Method for the Discrimination of Neutrons and gamma Rays With Organic Scintillation Detectors Using Frequency Gradient Analysis. *IEEE Transactions on Nuclear Science*, *57*(3), 1682–1691. https://doi.org/10.1109/TNS.2010.2044246

Liu, H. W., & Wang, G. J. (2007). Multi-criteria decision-making methods based on intuitionistic fuzzy sets. *European Journal of Operational Research*, *179*(1), 220–233. https://doi.org/10.1016/j.ejor.2006.04.009

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. *2010 IEEE International Conference on Data Mining*, 911–

916. https://doi.org/10.1109/ICDM.2010.35

Lkhagva, B., Suzuki, Y. u., & Kawagoe, K. (2006). New time series data representation ESAX for financial applications. *22nd International Conference on Data Engineering Workshops, 2006.*, 17–22. https://doi.org/10.1109/ICDEW.2006.99

Maharaj, E. A., & D'Urso, P. (2011). Fuzzy clustering of time series in the frequency domain. *Information Sciences*, *181*(7), 1187–1211. https://doi.org/10.1016/j.ins.2010.11.031

Manning, C. D., Raghavan, P., & Schutze, H. (2008). Introduction to information retrieval. In *Cambridge University Press*. Cambridge University Press Cambridge.

Mao, J., & Jain, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Transactions on Neural Networks*, *7*(1), 16–29. https://doi.org/10.1109/72.478389

Marrone, S., Cano-Ott, D., Colonna, N., Domingo, C., Gramegna, F., Gonzalez, E. . M., … Wisshak, K. (2002). Pulse shape analysis of liquid scintillators for neutron studies. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *490*(1), 299–307. https://doi.org/10.1016/S0168-9002(02)01063-X

Masulli, F., & Rovetta, S. (2006). Soft transition from probabilistic to possibilistic fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, *14*(4), 516–527. https://doi.org/10.1109/TFUZZ.2006.876740

McBeth, G. W., Lutkin, J. E., & Winyard, R. A. (1971). A simple zero crossing pulse

shape discrimination system. *Nuclear Instruments and Methods*, *93*(1), 99–102. https://doi.org/10.1016/0029-554X(71)90144-3

Mecca, G., Raunich, S., & Pappalardo, A. (2007). A new algorithm for clustering search results. *Data & Knowledge Engineering*, *62*(3), 504–522. https://doi.org/10.1016/J.DATAK.2006.10.006

Méger, N., Rigotti, C., Pothier, C., Nguyen, T., Lodge, F., Gueguen, L., … Datcu, M. (2019). Ranking evolution maps for Satellite Image Time Series exploration: application to crustal deformation and environmental monitoring. *Data Mining and Knowledge Discovery*, *33*(1), 131–167. https://doi.org/10.1007/s10618-018-0591-9

Meier, A., Pedrycz, W., & Portmann, E. (2019). *Applying Fuzzy Logic for the Digital Economy and Society* (A. Meier, E. Portmann, & L. Terán, Eds.). https://doi.org/10.1007/978-3-030-03368-2

Mendel, J. M. (2007). Advances in type-2 fuzzy sets and systems. *Information Sciences*, *177*(1), 84–110. https://doi.org/10.1016/j.ins.2006.05.003

Miller, W. H. (2017). Radiation Detection and Measurement, 2nd Edition. *Nuclear Technology*, *90*(2), 266–266. https://doi.org/10.13182/nt90-a34420

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis*, *21*(4), 441–458.

Mishra, S., & Chawla, M. (2019). *A Comparative Study of Local Outlier Factor Algorithms for Outliers Detection in Data Streams*. https://doi.org/10.1007/978-

981-13-1498-8_31

Mitchell, H. B. (2003). On the Dengfeng - Chuntian similarity measure and its application to pattern recognition. *Pattern Recognition Letters*, *24*(16), 3101–3104. https://doi.org/10.1016/S0167-8655(03)00169-7

Mitchell, H. B. (2004). A correlation coefficient for intuitionistic fuzzy sets. *International Journal of Intelligent Systems*, *19*(5), 483–490. https://doi.org/10.1002/int.20004

Mitchell, H. B. B. (2005). Pattern recognition using type-II fuzzy sets. *Information Sciences*, *170*(2), 409–418. Retrieved from https://www.sciencedirect.com/science/article/pii/S0020025504000933

Möller-Levet, C. S., Klawonn, F., Cho, K.-H., & Wolkenhauer, O. (n.d.). Fuzzy Clustering of Short Time-Series and Unevenly Distributed Sampling Points. In *Advances in Intelligent Data Analysis V* (pp. 330–340). https://doi.org/10.1007/978-3-540-45231-7_31

Mörchen, F., Ultsch, A., & Hoos, O. (2005). Extracting interpretable muscle activation patterns with time series knowledge mining. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, *9*(3), 197–208. https://doi.org/10.3233/KES-2005-9304

Moszyński, M., Costa, G. J., Guillaume, G., Heusch, B., Huck, A., & Mouatassim, S. (1994). Study of n-γ discrimination with NE213 and BC501A liquid scintillators of different size. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *350*(1–2), 226–234. https://doi.org/10.1016/0168-9002(94)91169-X

Nakhostin, M. (2012). Recursive algorithms for digital implementation of neutron/gamma discrimination in liquid scintillation detectors. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *672*, 1–5. https://doi.org/10.1016/j.nima.2011.12.113

Narayanamoorthy, S., Geetha, S., Rakkiyappan, R., & Joo, Y. H. (2019). Interval-valued intuitionistic hesitant fuzzy entropy based VIKOR method for industrial robots selection. *Expert Systems with Applications*, *121*, 28–37. https://doi.org/10.1016/J.ESWA.2018.12.015

Nie, F., & Zhang, P. (2013). Fuzzy partition and correlation for image segmentation with differential evolution. *IAENG International Journal of Computer Science*, *40*(3), 164–172. Retrieved from http://www.iaeng.org/IJCS/issues_v40/issue_3/IJCS_40_3_03.pdf

Niennattrakul, V., Srisai, D., & Ratanamahatana, C. A. (2012). Shape-based template matching for time series data. *Knowledge-Based Systems*, *26*, 1–8. https://doi.org/10.1016/j.knosys.2011.04.015

Nunthanid, P., Niennattrakul, V., & Ratanamahatana, C. A. (2011). Discovery of variable length time series motif. *The 8th Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand - Conference 2011*, 472–475. https://doi.org/10.1109/ECTICON.2011.5947877

Ohana-Levi, N., Paz-Kagan, T., Panov, N., Peeters, A., Tsoar, A., & Karnieli, A. (2019). Time series analysis of vegetation-cover response to environmental factors and

residential development in a dryland region. *GIScience & Remote Sensing*, *56*(3), 362–387. https://doi.org/10.1080/15481603.2018.1519093

Pal, N., Pal, K., Keller, J. M., & Bezdek, J. C. (2005). A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, *13*(4), 517–530. https://doi.org/10.1109/TFUZZ.2004.840099

Pal, N. R., Pal, K., & Bezdek, J. C. (1997). A mixed c-means clustering model. *Proceedings of 6th International Fuzzy Systems Conference*, *1*, 11–21. https://doi.org/10.1109/FUZZY.1997.616338

Pankowska, A., & Wygralak, M. (2006). General IF-sets with triangular norms and their applications to group decision making. *Information Sciences*, *176*(18), 2713–2754. https://doi.org/10.1016/j.ins.2005.11.011

Perlibakas, V. (2004). Distance measures for PCA-based face recognition. *Pattern Recognition Letters*, *25*(6), 711–724. https://doi.org/10.1016/j.patrec.2004.01.011

Qamar, U. (2014). A dissimilarity measure based Fuzzy c-means (FCM) clustering algorithm. *Journal of Intelligent and Fuzzy Systems*, *26*(1), 229–238. https://doi.org/10.3233/IFS-120730

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356

Rani, S., & Sikka, G. (2012). Recent Techniques of Clustering of Time Series Data: A Survey. *International Journal of Computer Applications*, *52*(15), 1–9.

https://doi.org/10.5120/8282-1278

Rao, M. R. (1971). Cluster analysis and mathematical programming. *Journal of the American Statistical Association*, *66*(335), 622–626. https://doi.org/10.1080/01621459.1971.10482319

Ratanamahatana, C., & Keogh, E. (2005). Three myths about dynamic time warping data mining. *International Conference on Data Mining (SDM'05)*, 506–510.

Ratanamahatana, C., Keogh, E., Bagnall, A. J., & Lonardi, S. (2005). A Novel Bit Level Time Series Representation with Implications for Similarity Search and Clustering. *In Proc. 9th Pacific-Asian Int. Conf. on Knowledge Discovery and Data Mining (PAKDD'05)*, 771–777. Springer.

Ravi, V., Srinivas, E. R., & Kasabov, N. K. (2008). On-line evolving fuzzy clustering. *Proceedings - International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 2007*, *1*, 347–351. https://doi.org/10.1109/ICCIMA.2007.314

Reinert, G., Schbath, S., & Waterman, M. S. (2000). Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, *7*(1–2), 1–46.

Rhee, F. C. H., & Hwang, C. (2001). A type-2 fuzzy C-means clustering algorithm. *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, 1926–1929. https://doi.org/10.1109/NAFIPS.2001.944361

Rodríguez-Fernández, V., Menéndez, H. D., & Camacho, D. (2017). Analysing temporal performance profiles of UAV operators using time series clustering. *Expert Systems*

with Applications, *70*, 103–118. https://doi.org/10.1016/j.eswa.2016.10.044

Rohlf, F. J. (1974). Methods of Comparing Classifications. *Annual Review of Ecology and Systematics*, *5*(1), 101–113. https://doi.org/10.1146/annurev.es.05.110174.000533

Rokch, L., & Maion, O. (2005). Clustering Methods. In *Data Mining and Knowledge Discovery Handbook*. https://doi.org/10.1007/0-387-25465-X

Ronchi, E., Söderström, P. A., Nyberg, J., Andersson Sundén, E., Conroy, S., Ericsson, G., … Weiszflog, M. (2009). An artificial neural network based neutron–gamma discrimination and pile-up rejection framework for the BC-501 liquid scintillation detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *610*(2), 534–539. https://doi.org/10.1016/j.nima.2009.08.064

Roush, M. L., Wilson, M. A., & Hornyak, W. F. (1964). Pulse shape discrimination. *Nuclear Instruments and Methods*, *31*(1), 112–124. https://doi.org/10.1016/0029-554X(64)90333-7

Saadati, R., & Park, J. H. (2006). On the intuitionistic fuzzy topological spaces. *Chaos, Solitons and Fractals*, *27*(2), 331–344. https://doi.org/10.1016/j.chaos.2005.03.019

Saini, S., & Rani, P. (2017). A Survey on STING and CLIQUE Grid Based Clustering Methods. *International Journal of Advanced Research in Computer Science*, *8*(5), 2015–2017.

Sanderson, T. S., Scott, C. D., Flaska, M., Polack, J. K., & Pozzi, S. A. (2012). Machine

learning for digital pulse shape discrimination. *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, 199–202. https://doi.org/10.1109/NSSMIC.2012.6551092

Santiago, L. M., Bagán, H., Tarancón, A., & Garcia, J. F. (2014). Synthesis of plastic scintillation microspheres: Alpha/beta discrimination. *Applied Radiation and Isotopes*, *93*, 18–28. https://doi.org/10.1016/J.APRADISO.2014.04.002

Santos, J. M., & Embrechts, M. (2009). On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5769, pp. 175–184). https://doi.org/10.1007/978-3-642-04277-5_18

Savran, D., Löher, B., Miklavec, M., & Vencelj, M. (2010). Pulse shape classification in liquid scintillators using the fuzzy c-means algorithm. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *624*(3), 675–683. https://doi.org/10.1016/j.nima.2010.09.130

Sayal, R., & Kumar, V. V. (2011). A Novel Similarity Measure for Clustering Categorical Data Sets. *International Journal of Computer Application*, *17*(1), 25–30.

Schlotzhauer, S. (2007). *Elementary statistics using JMP*. Retrieved from https://books.google.com/books?hl=en&lr=&id=5JYM1WxGDz8C&oi=fnd&pg=PR3&dq=Elementary+Statistics+Using+JMP&ots=MZOht9zZOP&sig=IFCsAn4Nd9clwioPf3qS_QXPzKc

Shan, Q., Chu, S., Ling, Y., Cai, P., & Jia, W. (2016). Designing a new type of neutron detector for neutron and gamma-ray discrimination via GEANT4. *Applied Radiation and Isotopes*, *110*, 200–204. https://doi.org/10.1016/j.apradiso.2016.01.024

Shastri, A. A., Ahuja, K., Ratnaparkhe, M. B., Shah, A., Gagrani, A., & Lal, A. (2019). Vector Quantized Spectral Clustering Applied to Whole Genome Sequences of Plants. *Evolutionary Bioinformatics*, *15*, 117693431983699. https://doi.org/10.1177/1176934319836997

Shieh, J., & Keogh, E. (2008). i SAX. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, *KDD '08*, 623. https://doi.org/10.1145/1401890.1401966

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. *PLOS ONE*, *10*(12), e0144059. https://doi.org/10.1371/journal.pone.0144059

Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., & Herawan, T. (2014). Big Data Clustering: A Review. In *Computational Science and Its Applications – ICCSA 2014* (pp. 707–720). https://doi.org/10.1007/978-3-319-09156-3_49

Shirkhorshidi, A. S., Ying Wah, T., Shirkhorshidi, S. M. R., & Aghabozorgi, S. (2019). Evolving Fuzzy Clustering Approach (EFCA): An Epoch Clustering That Enables Heuristic Post Pruning. *IEEE Transactions on Fuzzy Systems*, 1–1. https://doi.org/10.1109/TFUZZ.2019.2956900

Shubair, A., Ramadass, S., & Altyeb, A. A. (2014). kENFIS: kNN-based evolving neuro-fuzzy inference system for computer worms detection. *Journal of Intelligent &*

Fuzzy Systems, *26*(4), 1893–1908. https://doi.org/10.3233/IFS-130868

Shukla, A. K., & Muhuri, P. K. (2019). Big-data clustering with interval type-2 fuzzy uncertainty modeling in gene expression datasets. *Engineering Applications of Artificial Intelligence*, *77*(September 2018), 268–282. https://doi.org/10.1016/j.engappai.2018.09.002

Sivarathri, S., & Govardhan, A. (2014). Analysis of Clustering Approaches for Data Mining In Large Data Sources. *International Journal on Recent and Innovation Trends in Computing and Communication*, *2*(9), 2590–2595. Retrieved from http://www.ijritcc.org

Söderström, P.-A., Jaworski, G., Valiente Dobón, J. J., Nyberg, J., Agramunt, J., de Angelis, G., … Wadsworth, R. (2019). Neutron detection and γ-ray suppression using artificial neural networks with the liquid scintillators BC-501A and BC-537. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *916*, 238–245. https://doi.org/10.1016/J.NIMA.2018.11.122

Söderström, P.-A., Nyberg, J., & Wolters, R. (2008). Digital pulse-shape discrimination of fast neutrons and rays. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *594*(1), 79–89. https://doi.org/10.1016/j.nima.2008.06.004

Sosa, C. S., Flaska, M., & Pozzi, S. A. (2016). Comparison of analog and digital pulse-shape-discrimination systems. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated*

Equipment, *826*, 72–79. https://doi.org/10.1016/j.nima.2016.03.088

Sperr, P., Spieler, H., Maier, M. R., & Evers, D. (1974). A simple pulse-shape discrimination circuit. *Nuclear Instruments and Methods*, *116*(1), 55–59. https://doi.org/10.1016/0029-554X(74)90578-3

Steinley, D. (2004). Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods*, *9*(3), 386.

Studholme, C., Hill, D. L. G., & Hawkes, D. J. (1999). An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, *32*(1), 71–86.

Szmidt, E., & Kacprzyk, J. (2002). Using intuitionistic fuzzy sets in group decision making. *Control and Cybernetics*, *31*(4), 1037–1053. Retrieved from http://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-article-BAT2-0001-1826

Szmidt, E., & Kacprzyk, J. (2003). A consensus-reaching process under intuitionistic fuzzy preference relations. *International Journal of Intelligent Systems*, *18*(7), 837–852. https://doi.org/10.1002/int.10119

Szmidt, E., & Kacprzyk, J. (2004). A Similarity Measure for Intuitionistic Fuzzy Sets and Its Application in Supporting Medical Diagnostic Reasoning. In *Artificial Intelligence and Soft Computing - ICAISC 2004. ICAISC 2004. Lecture Notes in Computer Science* (pp. 388–393). https://doi.org/10.1007/978-3-540-24844-6_56

Tasdemir, K., & Merényi, E. (2011). A validity index for prototype-based clustering of data sets with complex cluster structures. *IEEE Transactions on Systems, Man, and*

Cybernetics, Part B: Cybernetics, *41*(4), 1039–1053. https://doi.org/10.1109/TSMCB.2010.2104319

Tsai, D.-M., & Lin, C.-C. (2011). Fuzzy C-means based clustering for linearly and nonlinearly separable data. *Pattern Recognition*, *44*(8), 1750–1760. https://doi.org/10.1016/J.PATCOG.2011.02.009

Uchida, Y., Takada, E., Fujisaki, A., Isobe, M., Ogawa, K., Shinohara, K., … Iguchi, T. (2014). A study on fast digital discrimination of neutron and gamma-ray for improvement neutron emission profile measurementa. *Review of Scientific Instruments*, *85*(11), 11E118. https://doi.org/10.1063/1.4891711

Van Rijsbergen, C. J. (1979). *Information retrieval*. Butterworths, London Boston.

Veenman, C. J., Reinders, M. J. T., & Backer, E. (2002). A maximum variance cluster algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(9), 1273–1280. https://doi.org/10.1109/TPAMI.2002.1033218

Wang, H., Wang, W., Yang, J., & Philip, S. (2002). Clustering by pattern similarity in large data sets. *2002 ACM SIGMOD International Conference on Management of Data*, *2*, 394–405. https://doi.org/10.1145/564691.564737

Wang, N. Y., & Chen, S. M. (2009). Temperature prediction and TAIFEX forecasting based on automatic clustering techniques and two-factors high-order fuzzy time series. *Expert Systems with Applications*, *36*(2 PART 1), 2143–2154. https://doi.org/10.1016/j.eswa.2007.12.013

Wang, W., & Xin, X. (2005). Distance measure between intuitionistic fuzzy sets. *Pattern*

Recognition Letters, *26*(13), 2063–2069. https://doi.org/10.1016/j.patrec.2005.03.018

Wang, X., Smith, K., & Hyndman, R. (2006). Characteristic-Based Clustering for Time Series Data. *Data Mining and Knowledge Discovery*, *13*(3), 335–364. https://doi.org/10.1007/s10618-005-0039-x

Weng, F., Jiang, Q., Chen, L., & Hong, Z. (2007). Clustering Ensemble based on the Fuzzy KNN Algorithm. *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007)*, 1001–1006. https://doi.org/10.1109/SNPD.2007.504

Wilson, D. R., & Martinez, T. R. (1997). Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, *6*, 1–34. https://doi.org/10.1613/jair.346

Wolski, D., Moszyński, M., Ludziejewski, T., Johnson, A., Klamra, W., & Skeppstedt, Ö. (1995). Comparison of n-γ discrimination by zero-crossing and digital charge comparison methods. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *360*(3), 584–592. https://doi.org/10.1016/0168-9002(95)00037-2

Wu, Y.-X., Han, X., Chen, C., Zou, L.-X., Dong, Z.-C., Zhang, Y.-L., & Li, H.-H. (2019). Time Series Gene Expression Profiling and Temporal Regulatory Pathway Analysis of Angiotensin II Induced Atrial Fibrillation in Mice. *Frontiers in Physiology*, *10*, 597. https://doi.org/10.3389/fphys.2019.00597

Xie, Z., Wang, S., & Chung, F. L. (2008). An enhanced possibilistic C-Means clustering algorithm EPCM. *Soft Computing*, *12*(6), 593–611. https://doi.org/10.1007/s00500-

007-0231-6

Xiong, H., Wu, J., & Chen, J. (2009). K-means clustering versus validation measures: a data-distribution perspective. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions On*, *39*(2), 318–331. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4711107

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141

Xu, Z. (2006a). A survey of preference relations. *International Journal of General Systems*, *36*(2), 179–203. https://doi.org/10.1080/03081070600913726

Xu, Z. (2006b). *On Correlation Measures of Intuitionistic Fuzzy Sets*. https://doi.org/10.1007/11875581_2

Xu, Z. (2007a). Intuitionistic Fuzzy Aggregation Operators. *IEEE Transactions on Fuzzy Systems*, *15*(6), 1179–1187. https://doi.org/10.1109/TFUZZ.2006.890678

Xu, Z. (2007b). Intuitionistic preference relations and their application in group decision making☆. *Information Sciences*, *177*(11), 2363–2379. https://doi.org/10.1016/j.ins.2006.12.019

Xu, Z. (2007c). Models for multiple attribute decision making with intuitionistic fuzzy information. *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, *15*(3), 285–297. https://doi.org/10.1142/S0218488507004686

Xu, Z. (2007d). Some similarity measures of intuitionistic fuzzy sets and their applications to multiple attribute decision making. *Fuzzy Optimization and Decision*

Making, *6*(2), 109–121. https://doi.org/10.1007/s10700-007-9004-z

Xu, Z., Chen, J., & Wu, J. (2008). Clustering algorithm for intuitionistic fuzzy sets. *Information Sciences*, *178*(19), 3775–3790. https://doi.org/10.1016/J.INS.2008.06.008

Xu, Z., & Yager, R. R. (2006). Some geometric aggregation operators based on intuitionistic fuzzy sets. *International Journal of General Systems*, *35*(4), 417–433. https://doi.org/10.1080/03081070600574353

Yanagida, T., Watanabe, K., & Fujimoto, Y. (2015). Comparative study of neutron and gamma-ray pulse shape discrimination of anthracene, stilbene, and p-terphenyl. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *784*, 111–114. https://doi.org/10.1016/j.nima.2014.12.031

Yanagida, T., Watanabe, K., Okada, G., & Kawaguchi, N. (2019). Neutron and gamma-ray pulse shape discrimination of LiAlO2 and LiGaO2 crystals. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *919*, 64–67. https://doi.org/10.1016/J.NIMA.2018.11.135

Yang, L., Lv, H., & Wang, W. (2006). Soft Cluster Ensemble Based on Fuzzy Similalrity Measure. *The Proceedings of the Multiconference on "Computational Engineering in Systems Applications,"* 1994–1997. https://doi.org/10.1109/CESA.2006.4281966

Yao, L., & Weng, K. (2012). On A Type-2 Fuzzy Clustering Algorithm. *The Fourth International Conferences on Pervasive Patterns and Applications*, 45–50.

Retrieved from http://www.thinkmind.org/index.php?view=article&articleid=patterns_2012_2_30_70033

Ye, F., Luo, W., Dong, M., He, H., & Min, W. (2019). SAR Image Retrieval Based on Unsupervised Domain Adaptation and Clustering. *IEEE Geoscience and Remote Sensing Letters*, *16*(9), 1482–1486. https://doi.org/10.1109/lgrs.2019.2896948

Yeung, K., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, *17*(10), 977–987.

Yeung, K. K., Haynor, D. D., & Ruzzo, W. (2001). Validating clustering for gene expression data. *Bioinformatics*, *17*(4), 309–318.

Yildiz, N., & Akkoyun, S. (2013). Neural network consistent empirical physical formula construction for neutron–gamma discrimination in gamma ray tracking. *Annals of Nuclear Energy*, *51*, 10–17. https://doi.org/10.1016/j.anucene.2012.07.042

Yousefi, S., Lucchese, L., & Aspinall, M. D. (2009). Digital discrimination of neutrons and gamma-rays in liquid scintillators using wavelets. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *598*(2), 551–555. https://doi.org/10.1016/j.nima.2008.09.028

Yu, X., Zhu, J., Lin, S., Wang, L., Xing, H., Zhang, C., … Tang, C. (2015). Neutron–gamma discrimination based on the support vector machine method. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators,*

Spectrometers, Detectors and Associated Equipment, *777*, 80–84. https://doi.org/10.1016/j.nima.2014.12.087

Zadeh. (1965). Fuzzy sets. *Information and Control*, (8), 338–353. Retrieved from https://www-liphy.ujf-grenoble.fr/pagesperso/bahram/biblio/Zadeh_FuzzySetTheory_1965.pdf

Zadeh, L. . (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, *1*(1), 3–28. https://doi.org/10.1016/0165-0114(78)90029-5

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning-I. *Information Sciences*, *8*(3), 199–249. https://doi.org/10.1016/0020-0255(75)90036-5

Zadeh, L. A. (1979). Approximate Reasoning Based on Fuzzy Logic. *IJCAI'79: Proceedings of the 6th International Joint Conference on Artificial Intelligence*, 1004–1010. Retrieved from https://dl.acm.org/citation.cfm?id=1623140

Zahn, C. T. (1971). Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers*, *C–20*(1), 68–86. https://doi.org/10.1109/T-C.1971.223083

Zakaria, J., Mueen, A., Keogh, E., & Young, N. (2016). Accelerating the discovery of unsupervised-shapelets. *Data Mining and Knowledge Discovery*, *30*(1), 243–281. https://doi.org/10.1007/s10618-015-0411-4

Zeng, J., Xie, L., & Liu, Z.-Q. (2008). Type-2 fuzzy Gaussian mixture models. *Pattern Recognition*, *41*, 3636–3643. https://doi.org/10.1016/j.patcog.2008.06.006

Zhang, H., Ho, T. B., Zhang, Y., & Lin, M. S. (2006). Unsupervised Feature Extraction for Time Series Clustering Using Orthogonal Wavelet Transform. *Informatica*, *30*(3), 305–319.

Zhang, J. S., & Leung, Y. W. (2004). Improved possibilistic C-means clustering algorithms. *IEEE Transactions on Fuzzy Systems*, *12*(2), 209–217. https://doi.org/10.1109/TFUZZ.2004.825079

Zhang, Q., & Chen, Z. (2014). A distributed weighted possibilistic c-means algorithm for clustering incomplete big sensor data. *International Journal of Distributed Sensor Networks*, *2014*(5), 430814. https://doi.org/10.1155/2014/430814

Zhao, Y., & Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, *55*(3), 311–331. Retrieved from http://link.springer.com/article/10.1023/B:MACH.0000027785.44527.d6