A PARTITION BASED FEATURE SELECTION APPROACH FOR MIXED DATA CLUSTERING

ASHISH DUTT

FACULTY OF COMPUTER SCIENCE
AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

A PARTITION BASED FEATURE SELECTION APPROACH FOR MIXED DATA CLUSTERING

ASHISH DUTT

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

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Ashish Dutt

Matric No: WHA130067

Name of Degree: Doctor of Philosophy (PhD)

Title of Thesis ("this Work"): A Partition based feature selection approach for mixed data clustering

Field of Study: Computer Science

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 copyright 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 copyright work;
(5) I hereby assign all and every rights 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 Ashish Dutt Date: 27/04/2020

Subscribed and solemnly declared before,

Witness's Signature Date: 27/4/2020

Name:

Designation:

A PARTITION BASED FEATURE SELECTION APPROACH FOR MIXED DATA CLUSTERING

ABSTRACT

Presently, educational institutions compile and store huge volumes of data, such as student enrolment and attendance records, as well as their examination results. Mining such data yields stimulating information that serves its handlers well. Rapid growth in educational data points to the fact that distilling massive amounts of data requires a more sophisticated set of algorithms. This issue led to the emergence of the field of Educational Data Mining (EDM). Traditional data mining algorithms cannot be directly applied to educational problems, as they may have a specific objective and function. This implies that a pre-processing algorithm has to be enforced first and only then some specific data mining methods can be applied to the problems. One such pre-processing algorithm in EDM is clustering. It is a widely used method in data mining to discover unique patterns in underlying data. It finds patterns by analysing the features in data. A feature contains a measured value. A value can be of an atomic type like categorical (text only) or numerical (number only). A categorical data type can be ordinal (ordered) or nominal (unordered). In either case, the feature is of univariate data type. Often in real-world environment, data consist of both categorical and numerical valued features. Such datasets are called mixed data. In literature, several clustering methods exist for analysing numerical or categorical data. There are a few clustering algorithms for handling mixed data. Clustering mixed data is dependent on the dissimilarities of its constituent features. This dependability on data types may influence a clustering solution. Assigning appropriate weights to the feature, such that it diminishes the data type influence may improve the performance of a partition clustering algorithm. In this thesis, a novel weighted feature selection approach on nominal features is proposed, for a partition

clustering algorithm that can handle mixed data. The proposed approach exploits the pre-processing nature of the partition clustering algorithm in the selection of weight assignment for nominal features. The benefits of weighting are demonstrated on both simulated and real-world mixed datasets. The experimental results yield better results for weighted nominal features in mixed data clustering.

Keywords- clustering, educational data mining, mixed data

.

PENDEKATAN PEMBAHAGIAN BERDASARKAN PEMILIHAN CIRI UNTUK PENGKLASTERAN DATA CAMPURAN

ABSTRAK

Pada masa kini, institusi pendidikan menyimpan dan menyusun data pada jumlah yang besar. Data ini termasuklah rekod pendaftaran, kehadiran, dan keputusan peperiksaan pelajar. Dengan menjalankan perlombongan data, hasilnya dapat merangsang maklumat bermanfaat kepada pengendali. Pertumbuhan pesat dalam data pendidikan juga menunjukkan bahawa penyulingan data secara besar-besaran memerlukan set algoritma yang lebih canggih. Hal ini membawa kepada kemunculan bidang Perlombongan Data Pendidikan (EDM). Algoritma perlombongan data yang tradisional tidak boleh diterapkan secara terus kepada permasalahan pendidikan. Hal ini kerana, algoritma sedia ada mungkin mempunyai objektif dan fungsi tertentu. Oleh itu, algoritma pra-proses, perlu diselaras terlebih dahulu dan cuma terdapat beberapa kaedah perlombongan data yang boleh diguna pakai bagi permasalahan tertentu. Salah satu kaedah dalam EDM ialah pengklasteran. Kaedah ini telah digunakan secara meluas dalam perlombongan data bagi menentukan corak unik pada data dengan menganalisis ciri-ciri tertentu. Ciri-ciri ini mengandungi nilai yang boleh diukur, sama ada nilai atomik seperti kategori (teks sahaja) atau berterusan (nombor sahaja). Jenis data juga boleh dikategorikan kepada ordinal (teratur) atau nominal (tidak teratur). Hakikatnya, dalam dunia nyata, data boleh terdiri daripada yang nilainya bersifat kategori dan berterusan. Data jenis data ini dikenali sebagai data campuran. Dalam literatur, terdapat beberapa kaedah pengklasteran untuk data campuran dan ia bergantung kepada ketidaksamaan unsur pada ciri-ciri data. Kebergantungan terhadap jenis data ini boleh mempengaruhi penyesuaian pengklasteran. Dengan menggunakan wajaran yang sesuai, ia boleh mengurangkan pengaruh jenis data, seterusnya mempertingkatan prestasi algoritma pembahagian pengklasteran. Dalam tesis

ini, pendekatan pemilihan ciri wajaran yang baharu pada ciri nominal adalah dicadangkan bagi algoritma pembahagian pengklasteran yang juga boleh mengendalikan data bercampur. Pendekatan yang dicadangkan ini dapat mengeksploitasi sifat pra-pemprosesan bagi algoritma pembahagian pengklasteran dalam pemilihan tugasan wajaran untuk ciri nominal. Faedah daripada wajaran ini juga dapat ditunjukkan daripada kedua-dua dataset, sama ada data daripada simulasi campuran atau data daripada dunia nyata. Dapatan eksperimen juga memberikan hasil yang lebih baik untuk ciri wajaran nominal dalam pengklasteran data bercampur.

Kata kunci- pengklasteran, perlombongan data pendidikan, data bercampur

ACKNOWELDGEMENTS

"The teacher who is indeed wise does not bid you to enter the house of his wisdom but rather leads you to the threshold of your mind",

Khalil Gibran.

I would like to begin by expressing sincere gratitude to Almighty for spearing my life with good health to witness the successful completion of my PhD research

Teaching is considered a noble profession and so are teachers. The Almighty has blessed me by giving his trusted advisors to be my guides. I would like to thank PhD supervisors, Assoc. Prof. Dr. Maizatul Akmar Ismail and Dr. Hoo Wai Lam whom despite their busy schedules, invested their valuable time towards shaping this journey towards a great learning experience. I would also like to thank Dr. Rashmi Gangwar, consultant at UNICEF who helped me in acquiring the DISE school panel level dataset from the Ministry of Education, India.

While a teacher shapes the young mind to be a responsible citizen, it's the parents, who are responsible for germination of that young mind by providing it with care, wisdom and above all a valued education. Words are scarce to convey my gratitude to my father and mother who have provided for me thus far, and have accorded the gift of education. To my younger brother who has always provided me with much needed moral support in times of despair, thank you.

This thesis is dedicated to my family, father Hari Dutt Gangwar, mother Kiran Gangwar, younger brother Rahul Dutt Gangwar, his wife Swapnil and our little bundle of joy, Samaira Dutt.

Kuala Lumpur, December 31, 2019 A.D

TABLE OF CONTENTS

ABSTRACT iv
ABSTRAK vi
ACKNOWELDGEMENTS viii
TABLE OF CONTENTS ix
LIST OF FIGURES xii
LIST OF TABLES xiii
LIST OF SYMBOLS AND ABBREVIATIONS xiv

CHAPTER 1: INTRODUCTION 1
1.1. Introduction 1
1.2. Overview 1
1.3. Background 2
1.4. Problem Statement 4
1.5. Research Aim & Objectives 7
1.6. Research Questions 7
1.7. Research Scope 7
1.8. Research Methodology 8
1.9. Research Significance 9
1.10. Thesis Organisation 9
1.11. Chapter Summary 10

CHAPTER 2: LITERATURE REVIEW 11
2.1. Introduction 11
2.2. Taxonomy for mixed data clustering 14
2.3. An overview on the types of data 15
2.4. An overview of mixed data unsupervised algorithms 16
2.4.1. Partition-based clustering 16
2.4.1.1. Cluster Centre Initialization 22
2.4.1.2. Number of Clusters 31
2.4.2. Hierarchical-based clustering 35
2.4.3. Model-based clustering 39
2.4.4. Neural network-based clustering 42
2.4.5. Key literature review observations 47

2.5. An overview of distance measures 48
2.5.1. Distance measures for numerical data 48
2.5.2. Distance measures for categorical data 50
2.5.2.1. Simple Matching Coefficient (SMC) 52
2.5.2.2. Inverse Occurrence Frequency (IOF) 53
2.5.2.3. Lin measure 53
2.5.3. Distance measures for mixed data 53
2.6. An overview of unsupervised feature selection approaches 57
2.7. An overview of data transformation approaches 61
2.7.1. Discretisation 61
2.7.2. Numerical coding 61
2.8. Cluster Validation Metric (CVM) 62
2.9. Educational Data Mining (EDM) 68
2.9.1. Clustering algorithms applied in EDM 68
2.9.2. Mixed data clustering approach in EDM 79
2.10. Chapter summary 80

CHAPTER 3: RESEARCH METHODOLOGY 82
3.1. Introduction 82
3.2. Research approach 82
3.3. Research methodology 84
3.3.1. Feature detection stage 84
3.3.2. Feature cleaning stage 85
3.3.2.1. Results Normalisation 85
3.3.3. Feature selection stage 86
3.3.3.1. Supervised Feature Selection (SFS) methods 86
3.3.3.2. Unsupervised Feature Selection (UFS) methods 86
3.3.4. Feature evaluation stage or Cluster validation metric 87
3.4. Chapter summary 88

CHAPTER 4: ALTERNATIVE APPROACH TO UNSUPERVISED FEATURE SELECTION for MIXED DATA CLUSTERING 89
4.1. Introduction 89
4.2. The K-prototypes algorithm for mixed data clustering by Huang (1997) 89
4.2.1. The K-prototypes algorithm drawbacks 90
4.3. An improved K-prototypes clustering algorithm for mixed data clustering by Ji et al. (2013) 91

4.3.1. Drawback of the improved K-prototypes algorithm 92
4.4. The Gower dissimilarity measure for mixed dataset by J.C. Gower (1971) 93
4.4.1. Drawback of the Gower Dissimilarity Measure 94
4.5. The Proposed Unsupervised Feature Selection Algorithm for Mixed Data Clustering (UFSMDC) 95
4.5.1. Preliminaries 95
4.5.2. Flow chart of the proposed approach 96
4.5.3. The proposed approach 98
4.5.4. The difference between proposed approach and Huang's approach (1997) 101
4.5.5. Comparison of the results 101
4.5.6. Performance analysis of the proposed approach 103
4.6. Summary of the chapter 105

CHAPTER 5: APPLICATION OF UFSMDC APPROACH ON MIXED DATASETS AND RESULT 106
5.1. Introduction 106
5.2. Baseline Methods 106
5.3. Performance Evaluation Metrics 106
5.4. Application of the proposed approach on mixed educational dataset 107
5.4.1. Experimental Setup 107
5.4.2. A brief description of other mixed datasets 114
5.4.3. Application of the proposed approach on other mixed datasets 114
5.5. Results Comparison and Discussion 116
5.6. Chapter Summary 120

CHAPTER 6: CONCLUSION AND FUTURE WORK 121
6.1. Introduction 121
6.2. Summary of Research Findings 121
6.3. Limitations of the Study 124
6.4. Recommendation for Future Work 124
REFERENCES 125
LIST OF PUBLICATIONS AND PAPERS PRESENTED 142

LIST OF FIGURES

Figure 2.1: Types of clustering algorithms……12

Figure 3.1: Proposed feature clustering approach…...……...83

Figure 4.1: Flowchart of the proposed approach…...…….....97

Figure 4.2: An unsupervised feature selection approach for mixed data clustering….100

Figure 4.3: Number of iterations performed by the two algorithms….……....102

Figure 5.1: Primary school enrolment and demographic distribution.....……....109

Figure 5.2: Primary school enrolment and school facilities distribution...……...109

Figure 5.3: Outliers in numerical features...…….111

Figure 5.4: Outliers in numerical features...…….111

Figure 5.5: Partial outlier treatment for numerical features..…….....112

Figure 5.6: Partial outlier treatment for numerical features..…….....112

Figure 5.7: Determining the number of clusters using silhouette width..……113

LIST OF TABLES

Table 2.1: Research on mixed data clustering…...……...15

Table 2.2: Summary of hierarchical clustering methods for mixed data…….....37

Table 2.3: Distance measures for numerical data…….....49

Table 2.4: Distance measures for categorical data…….....51

Table 2.5: Distance measures for mixed data…….....57

Table 2.6: Filter based unsupervised feature selection methods for mixed data clustering…….59

Table 2.7: Internal Cluster Validation Metrics……63

Table 2.8: Clustering methods adapted in EDM…….....69

Table 2.9: Mixed data clustering in EDM……80

Table 4.1: The K-prototypes algorithm by (Huang, 1997)…….....90

Table 4.2: An improved K-prototypes algorithm by (Ji et al., 2015)……..92

Table 4.3: The comparison results……104

Table 5.1: Statistical properties of the dataset……108

Table 5.2: Application of PAM, CLARA with and without proposed UFSMDC approach on mixed data…….....115

Table 5.3: Five times execution cycle of Gower coefficient in K-prototypes on school panel level dataset……...….....117

Table 5.4: Application of K-prototypes algorithm using the Gower coefficient on school panel level dataset……..117

Table 5.5: Five times execution cycle of modified Gower coefficient in K-prototypes on school panel level dataset……..118

Table 5.6: Application of K-prototypes algorithm using the modified Gower coefficient (UFSMDC) approach on school panel level dataset…….118

LIST OF SYMBOLS AND ABBREVIATIONS

ARI	:	Adjusted Rand Index
CCA	:	Canonical Correlation Analysis
CHI	:	Callinski-Harabaz Index
CVM	:	Cluster Validation Metric
DBI	:	Davies-Bouldin Index
DI	:	Dunn Index
DM	:	Data Mining
EDM	:	Educational Data Mining
FC	:	Feature Clustering
FE	:	Feature Extraction
FS	:	Feature Selection
KDD	:	Knowledge Discovery in Databases
LDA	:	Linear Discriminant Analysis
PCA	:	Principal Component Analysis
RI	:	Rand Index
SC	:	Silhouette Coefficient
SFS	:	Supervised Feature Selection
UFS	:	Unsupervised Feature Selection
UFSMDC:		Unsupervised Feature Selection for Mixed Data Clustering

LIST OF APPENDICES

Appendix A : Modified Gower distance function …….. 143

Appendix B : Modified Gower distance usage ……... 145

Appendix C : A sample of 04 UCI ML datasets and the school panel level dataset used for experiments ……148

CHAPTER 1: INTRODUCTION

1.1. Introduction

This chapter presents an introduction to Educational Data Mining (EDM) and Mixed Data Clustering (MDC). It consists of the problem statement, research questions, the aim and objectives of this research, followed by the research scope as well as the research methodology and finally the significance of this work. It then describes the outline of the thesis.

1.2. Overview

The field of EDM, includes applications and methods aimed at understanding how learners learn, in specific reference to their environmental, socio-economic and even psychological conditions (Baker, 2010). The implementation of EDM in an educational environment is limited (Ranjan & Malik, 2007). In general, there have been several methods for the Data Mining (DM) process, but some researchers argue that these are too generic to be applied to a specialized context like EDM, (Baker & Yacef, 2009). It is said that universities across the globe now have the additional responsibility of ensuring successful students (Campbell, DeBlois, & Oblinger, 2007). With a wider acceptability of EDM as a branch of DM, there have been a plethora of research studies in educational areas such as pupil failure, pupil dropout rate and pupil low attendance (Jing, 2004). A researcher suggested a predictive algorithm as a method to curb the pupil school or course dropout problem (Lin, 2012). E-commerce websites use recommender system to suggest similar items to user's browsing their website. Similar approaches were applied to model the student behavioural pattern in educational context, but because of the highly domain dependency, they failed. (Santos & Boticario, 2010).

Clustering is an unsupervised method applied to coalesce data into groups such that similar features align together (Khan & Ahmad, 2013; Jain & Dubes, 1988). There are several clustering

algorithms for processing data in either the form of numerical or categorical feature values (Hall, Witten, & Frank, 2011). The clustering algorithms group the attributes on some idea of "similarity". To calculate similarity for numerical features, arithmetic calculations (like distance operations or mode), are computed. However, the same operation cannot be applied to categorical data which contains feature values that are not inherently ordered such as, "blue, black, and white". To calculate similarity dependent on distance for categorical features is a challenging task (Boriah, Chandola, & Kumar, 2008).

In practice a majority of the datasets comprise of both numeric and categorical feature-set. Such datasets are better known as mixed data. The mixed data is found in several application areas such as in education, health, marketing or financial institutions. (Ahmad & Dey, 2007), (Morlini & Zani, 2010).

1.3. Background

Poised to meet the growing requirement of pervasive student engagement, is the young field of EDM. The process of EDM is a multi-dimensional process involving data extraction from either offline educational environment like attendance records maintained in physical registers of schools or colleges, or, digitised educational records stored in educational information systems like a learning management system. Such datasets have unique characteristics, and when processed they yield invaluable information. Such information can be used to help both the educator and the student in improving their task performance. (Dutt, Aghabozrgi, Ismail, & Mahroeian, 2015). Traditionally researchers have applied either or both supervised and unsupervised data mining algorithms such as association rules, support vector machine, random forest, and logistic regression to datasets collected from educational environments. The researchers (Dutt, Ismail, & Herawan, 2017) covered a three-decade long range (1986-2016) of research on the applicability of unsupervised algorithms to educational data. Zaıane & Luo (2001) proposed that data mining methods can be used to determine student interaction patterns in e-learning courses. Zaıane (2002), Tang &

McCalla (2005) advised that the e-learning systems can be made effective by integrating DM methods like association rules and unsupervised learning in them. Yet another group of researchers Baker, Corbett, & Koedinger (2004) conducted an interesting case study on human and computer interaction. Their idea was to apply DM methods as tools for understanding how human can game the e-learning system. Hoppe (2003) suggested the integration of web based tools in e-learning environment in order to support EDM related activities by educators. Beck & Woolf (2000) suggested various methods related to classification methods applicable in EDM context. Some other lesser known modelling methods like student modelling are used too and is an emerging research discipline in EDM (Baker & Yacef, 2009). Another group of researchers developed a software to extract student text records to yield statistics, which was then integrated into a learning management system (García, Romero, Ventura, & de Castro, 2011). Although, the e-learning systems help educational institutions by providing an alternative learning system but the drawback is, they are not equipped with relevant tools. Which can be used to monitor, assess and provide feedback to student learning activities in using them. This essential lack of elements further constrains their usage and calls for additional external input (Zorrilla, Menasalvas, Marin, Mora, & Segovia, 2005). EDM is concerned with analysing data generated in an educational setup using disparate systems with the aim to develop models to improve learning experience and institutional effectiveness (Dutt et al., 2017).

The integration of data mining algorithms in the educational environment is a widely researched area. The underlying idea is to explore and analyse the educational datasets such that interesting patterns can be found. Such patterns can then help the educator, or the stakeholder or parents or even the students to improve upon their learning and teaching environments. Clustering is an imperative pre-processing algorithm that has widely been applied in EDM. Clustering is an unsupervised method aimed at understanding and processing data from an unsupervised perspective. It is widely used in diverse areas such as bioinformatics, pattern-recognition, statistics and machine learning (Dutt et al., 2017). The existing literature is replete with research works on either unsupervised

numerical or unsupervised categorical data. Thus clustering algorithms like K-means (Hartigan & Wong, 1979) and K-modes (Huang, 1998) were developed to analyse numerical or categorical data. With the passage of time, the complexity of datasets increased, hence now there are datasets with both categorical and numerical features occurring together (Hsu, Chen, & Su, 2007; Hsu & Huang, 2008). For example, a student report card will contain a feature like sex that is categorical and subject marks which are numerical. So, 'is the student report card dataset multivariate or mixed?' is a question from the naming perspective. The multivariate data as the term indicates is a dataset with several variables (Hair et al., 1998). The related concepts are univariate (one variable in a dataset) and bivariate (two variables in a dataset). In contrast, a mixed dataset is a dataset consisting of variables or features of different datatypes (Ahmad and Khan, 2019). Basis of this analogy, a student report card, when seen from a variable perspective will be deemed as multivariate. If the student report card is seen from the data type perspective therefore, it's a mixed dataset. This thesis emphasises on mixed data types. Basis on this, the question then arises is 'how to group-mix data type in an unsupervised fashion?'

1.4. Problem Statement

Feature Selection (FS), also known as dimensionality reduction or variable selection, is essentially a combinatorial optimization problem. It is a process to include variables with maximum variance from the original feature set, thereby eliminating redundant or less informative variables. The commonly known supervised FS methods are Fischer Score, Information Gain Relief, Chi Squares and Pearson correlation coefficients (Gu, Li, & Han, 2012) and (Weiss, 2015). Depending upon the availability of label information, feature selection is classified into supervised and unsupervised methods. If a class label is provided, then its supervised feature selection otherwise it's unsupervised. There also exists semi-supervised FS method that works, by integrating a small portion of labelled data into unlabelled data as an additional information, to improve the performance of Unsupervised Feature Selection (UFS) algorithms.

Often in literature and practice, FS is referred to as Feature Extraction (FE). However, there is a subtle difference between FS and FE. While the FS method aims to select a small subset of features that minimize redundancy and maximize relevance to the target such as class label in a classification task (Guyon & Elisseeff, 2003). In contrast, the FE method also known as feature transformation determines a weighted projection of several features into a new dimension and selects a predefined number of dimensions (Guyon & Elisseeff, 2006). The major drawback of FE methods is that the transformed variables (that are eigenvectors- transformed coefficients of each principal component) are difficult to map with the original variables (Wang, Lei, Zeng, Tong, & Yan, 2013). Some notable unsupervised FE methods include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Canonical Correlation Analysis (CCA) (Weiss, 2015).

A feature is an attribute containing data. Data is of several types such as text or categorical data, numerical data, image data and likewise. On the other hand, the type of data contained in a feature define its identity and properties. Often in a real-world environment, different types of features are observed interacting with each other. This thesis focusses on the educational environment. In an educational environment, there exist both categorical and numerical features which define an object such as a student or a teacher. An object is an abstract entity consisting of features. For instance, a student object S comprises of features such as subject marks containing numerical data or subject grades containing text data. It can be said, S consist of n numerical and m categorical features. Picking some notations from set theory, suppose a universe U (read school) consist of students S. Let S consist of features *subject-marks* (numeric n), *subject-grades* (categorical m). Let the subject marks consist of subjects such as English, Hindi, Science and Mathematics. They can be denoted as $n_1, n_2, \dots, n_n$. Now, suppose the subject grades are Good, Pass, Distinction and Fail. These can be denoted as $m_{1,}, m_2, \dots, m_n$ so the universe U is denoted as a set of features $U = \{n_1, n_2, \dots, n_n + m_1, m_2 + \cdots, m_n\}$. The feature distribution helps to determine the feature types

contained in a dataset and their distribution. Some well-known feature distributions are Gaussian or Normal distribution, Student's t-distribution and Chi-Squared distribution.

Educational institutions use information systems to process and store data like student attendance records or examination records. Such data can be either labelled or unlabelled. In literature there are many studies related to labelled or supervised FS in EDM (Sivakumar, Venkataraman, & Selvaraj, 2016),(Márquez-Vera et al., 2016),(Asif, Merceron, Ali, & Haider, 2017). Labelling the data is an expensive activity and this problem is compounded further when the data is unlabelled (Zhu, Zhang, Lin, & Shi, 2007),(Sheng, Provost, & Ipeirotis, 2008).

Hence the problem of this work is briefly stated as:

"An educational dataset consists of either or both categorical and numerical features. Applying conventional clustering algorithms like K-means or K-modes directly to such a dataset (hereafter known as mixed data) often leads to information loss, because they do not take into account the statistical properties like distribution of a feature."

There is negligible work on mixed data clustering in the context of EDM. Motivated from the actuality that a majority of data consist of both categorical and numerical features. This work contributes to scientific research by proposing a partition-based feature selection method for processing mixed data in EDM. The proposed method accounts for either or both categorical and numerical data types in a given dataset. The partition-based clustering methods are flexible methods that are based on iterative relocation of data points between the clusters, the extensive results have shown that the proposed approach yields high cluster purity when compared to existing partition-based clustering algorithms. It is essential to mention that clustering purity here refers to the internal information of the clustering process that evaluates the goodness of a clustering structure without reference to external information.

1.5. Research Aim & Objectives

The aim of this research is to propose a clustering approach for mixed datasets inherited in an educational environment to yield pure clusters. The objectives are:

4. To identify existing clustering approaches for treating mixed data in EDM.
5. To propose an alternative approach to unsupervised feature selection for mixed data clustering.
6. To evaluate the proposed approach with existing partition-based clustering methods for mixed datasets.

1.6. Research Questions

The research questions which are answered through this work are as follows;

Q1. What are the existing clustering algorithms that have been applied to mixed educational datasets?

Q2. Is FS a component of these existing clustering algorithms?

Q3. How much the proposed approach improves the purity of obtained clusters?

1.7. Research Scope

To ensure that this research can attain its defined objectives within a stipulated time frame, it is important to define the research scope as:

4. This research study focusses on primary schools of New Delhi, India, where primary is defined as grade 1 to grade 5. The schools are defined as either all boys only school, all girls only school or coeducational schools.
5. The focus of this work will be on the facilities provided by the schools (*categorical data*) and its impact on the student enrolment rate (*numerical data)*.

6. The research focus is on the cluster purity and its usefulness. A useful cluster is defined as a cluster that has captured the natural structure of the data. (Tan, Steinbach, & Kumar, 2013).

1.8. Research Methodology

A typical educational dataset consists of a combination of categorical and numerical features (*a mixed dataset*). A numerical feature like student exam score or teacher class hours will contain a numeric value. Similarly, a categorical feature like school location, will contain a text value. The proposed approach will accept educational data as an input and will yield cluster of similar features. In the first stage, the proposed approach will determine the nature of input data. It will then be segregated into categorical and numerical features. Both types of features will be treated for issues like missing data treatment, collinearity and multicollinearity, correlation, skewness, near zero variance and outliers. The idea is to conduct rigorous data pre-processing such that only the statistically relevant features remain in the data. It will also help in data dimensionality reduction. This subset is subjected to the proposed distance method for mixed data and the result is saved to a data matrix. Concurrently, the possible number of groups is determined by the Elbow method (Kaufman & Rousseeuw, 2009) and saved to a variable. By the end of first stage, a subset of the original features is obtained. This subset will contain only the important features and the subsequent clustering will become more efficient.

In the second stage, distance-based data matrix for mixed data and the number of possible groups from the first stage are passed into a partition-based clustering approach to obtain clusters.

In the third and final stage, the obtained clusters will be checked for cluster purity by using an internal clustering validation metric called the Silhouette Coefficient (SC), and the result will be evaluated with baseline methods.

1.9. Research Significance

The research questions in this work were inspired to address three interrelated topics that corresponded to the design and application of an operational clustering algorithm. This work uncovers the issues and opportunities in 1) analysing mixed data inherent in educational environment using an unsupervised approach, 2) identifying the best representative features, and 3) evaluating the performance of the unsupervised approach with baseline methods. Furthermore, this work also addresses the gap in existing literature on the applicability of partition-based clustering algorithm in EDM. Finally, this work used real life educational datasets for analysis, visualisation and validation of the proposed approach.

1.10. Thesis Organisation

This thesis consists of six chapters.

Chapter 1 presents an introduction to EDM, mixed data clustering and the motivation of the research work. Followed by the problem statement, research questions, research aim and objectives, a brief description of the contributions and significance of this research.

Chapter 2 presents a review on mixed data clustering algorithms. This chapter starts by presenting an overview of the different types of clustering algorithms. It particularly focuses on partition and hierarchical clustering algorithms. It also presents a review on existing studies focused on mixed data clustering. The definition of EDM is discussed, in particular reference to unsupervised algorithms in varied educational contexts. Then the strengths and weaknesses of the existing works are elaborated. This chapter provides a common platform which prompts further discussions over the next chapters.

Chapter 3 presents the methodology of the research.

Chapter 4 discusses the different experimental setups carried out for the implementation and validation of the proposed approach. It describes the dataset used, discloses the baseline methods, and explains the performance evaluation metrics.

Chapter 5 presents the experimental results of the different experiments carried out in the research. It also provides the results comparisons across the experiments and the other baseline methods. Chapter 6 discusses the research findings and compares them with other related studies. It concludes the research and shows the research contributions and limitations, as well as future research directions.

1.11. Chapter Summary

This chapter discusses the background behind this research work and defines the problem it intends to address. It outlines the research questions, aim and objectives of the study. It also summarizes the significance of this research work. The next chapter provides the fundamentals of clustering algorithms and its application to EDM.

CHAPTER 2: LITERATURE REVIEW

2.1. Introduction

Unsupervised learning refers to a branch of algorithms where the focus is on determining objects with similar properties or characteristics. The fundamental key to understanding the unsupervised learning also referred to as clustering is "similarity". The grouping of similar objects is also referred to as clusters. Within the unsupervised learning algorithms, clustering is the most widely used technique (Dutt et al., 2017). The size of the data can be reduced greatly, if similar data points are grouped together. This grouping can be achieved only through clustering. Once the groups are formed, they can be used for further analysis. But the researcher must be careful when defining what constituents as a group. The cluster definition plays an important role at this stage, for it determines not only the components of a group but also the component properties. In the absence of an adequate or improper cluster definition, there are strong possibilities of information loss.

Looking through literature, the categorization of clustering is not precise, as several categories concur with each other. In traditional computation terms, the clustering methods are overtly divided into two categories, namely, hierarchical and partition as shown in Figure 2.1. To understand clustering, its first imperative to comprehend supervised classification (or discriminant analysis). A supervised classification task require objects to be classified based on an assortment of pre-labelled objects (Kaufman & Rousseeuw, 2009). In contrast is the unsupervised classification or clustering, where the objects are asked to be classified in the absence of predefined labels (Kaufman & Rousseeuw, 2009). Thus, the process of unsupervised classification is much more difficult as compared to the supervised classification (Kaufman & Rousseeuw, 2009).

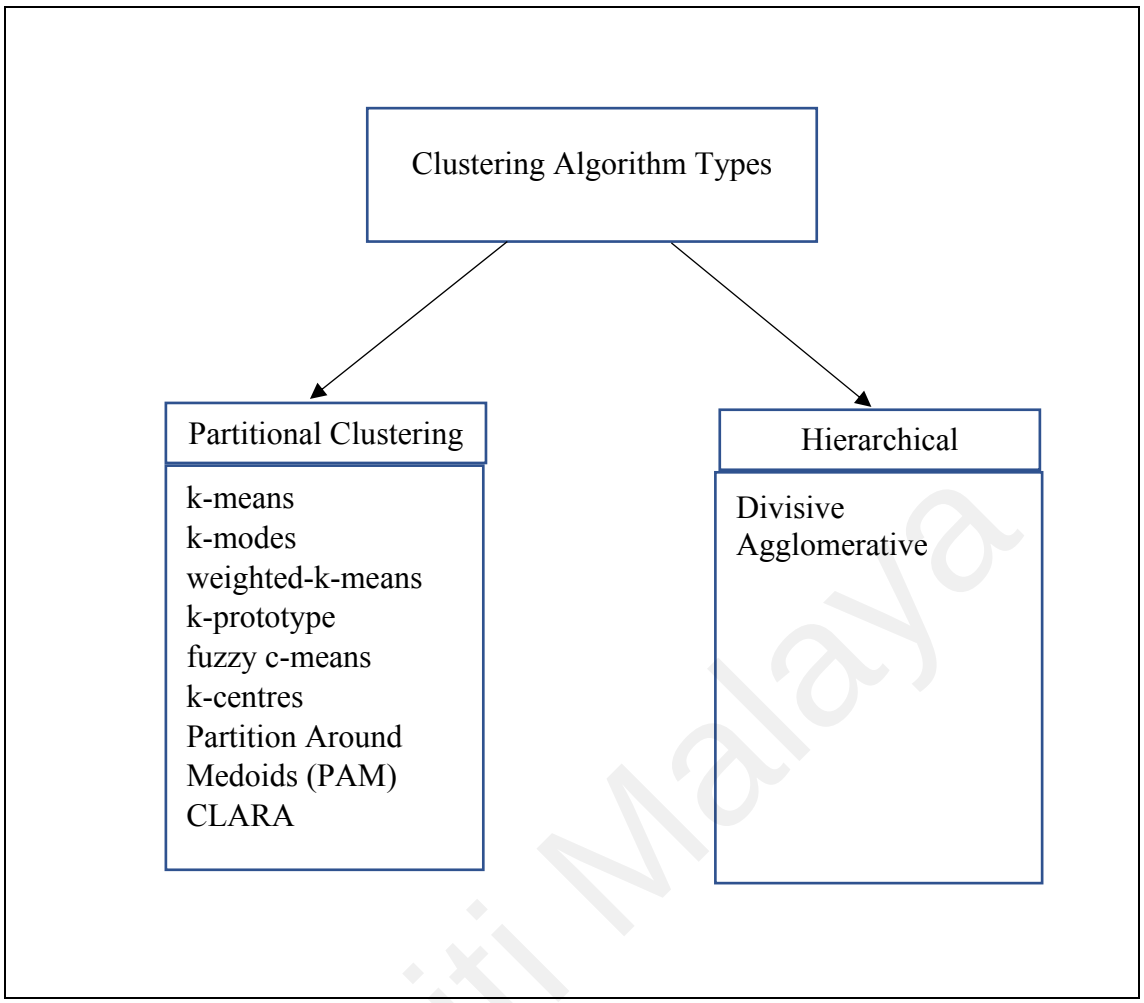

Figure 2.1: Types of clustering algorithms

When discussing the unsupervised learning methods, a few important points need to be taken into consideration. For instance, the cluster structure is an important component, for it can define the cluster to be a single or multi-layered clustering solution. On the basis of this layering structure, two major types of clustering methods are given, the partitional and the hierarchical algorithms. Other than this traditional clustering nomenclature, there are other clustering divisions that appear in the literature. For instance, the Hard Clustering and Soft Clustering algorithms, that were suggested as data points that are aligned to clusters (e.g., binary alignment vs the association degree), are in fact quite popular too (Pathak & Pal, 2016).

Some other clustering classifications are based on the algorithm implementation approach: for instance, methods dependent on data distribution centroid called the centroid-based clustering, methods dependent on data path networks called the grid-based clustering, or the methods dependent on data density called the density-based clustering and many others. And in literature, there is no dearth of algorithms that apply partition and or hierarchical or hard and or soft clustering algorithms to their respective use cases.

If the clustering definition is to be considered, then the available literature is ridden with algorithms like K-means/fuzzy c-means that are widely used algorithms for centroid dependent partitional clustering methods. Another notable example is SingleLink (SLINK) (Rodriguez et al., 2019), nearest-neighbour prominent hard hierarchical algorithm. Other clustering algorithms classify objects on the basis of their densities like the DBSCAN algorithm (Liu, Yang, & He, 2017).

Meanwhile, the other hierarchical algorithms that consider the objects geometrical structure also exist in literature like the Ward's method, which is a geometrical-based hierarchical clustering algorithm. Such methods are cluster centroid dependent. Yet another hierarchical clustering method is the Divisive clustering that takes the "top-down" approach. The idea behind this method is to begin the clustering process by amassing all the objects together into one big group. Thereafter split the big group into sub-groups dependent upon the object's hierarchy with each other. Further detail on these unsupervised methods can be found in Rodriguez et al (2019).

Because of the fundamental nature of clustering to either decompose the objects into cohesive structural units or recompose them, it is an important data pre-processing technique. And this method has been widely exploited in analysing voluminous data sizes like "big data". Some researchers have applied the big data clustering methods to students' examination records (Sagiroglu & Sinanc, 2013). Overtly speaking, an

education based information system is classified into two categories: Category one encompasses the traditional classrooms and the category two consist of the e-learning systems that include Learning Management Systems (LMSs) and the hypermedia systems (Parack, Zahid, & Merchant, 2012a) and the Intelligent Tutoring System (ITS) (Baker & Yacef, 2009).

The task of data clustering generally consist of two approaches. The first approach, *partition clustering*, creates an initial set of *K* partitions (clusters). In this approach, the number of partitions or *K* is required to be specified by the user. The second approach, *hierarchical clustering*, involved a structural tree concept that consist of either top-down or bottom-up clustering. Typically, in a data clustering problem, the objective is to find the cluster centres or the centroid. The partition-based clustering approach works by iteratively relocating the centroids (Rodriguez et al., 2019).

This work aims at grouping mixed data in an educational setting. It defines a mixed dataset as a collection of numerical (also called as continuous) and categorical (also called as nominal or ordinal) scaled variables based on the definition by Ahmad & Khan (2019).

2.2. Taxonomy for mixed data clustering

Recent years have seen an astounding growth in the number of clustering algorithms for data mining tasks. Most real-world data consist of both continuous and categorical data types also known as mixed data. Mixed data clustering can be applied in many a different way depending upon the data types. However, there exist no structured path for the research done in this area.

A taxonomy of the previous works is discussed in this sub-section. This taxonomy has identified three prominent research domains of unsupervised algorithms – *partitional,*

hierarchical and *density based*. In Table 2.1, highlights the suggested categorisation with three varied types of domains for grouping mixed data.

Table 2.1: Research on mixed data clustering

No.	Research theme	Research papers
1.	Partition approach for mixed data clustering	(Chiodi, 1990), (Huang, 1997; Huang, 1998), (Modha & Spangler, 2003), (Ahmad & Dey, 2007), (Zhao, Dai, & Tang, 2007), (Cheng & Leu, 2009), (Roy & Sharma, 2010), (Barcelo-Rico & Diez, 2012), (Liang, Zhao, Li, Cao, & Dang, 2012), (Ji, Bai, Zhou, Ma, & Wang, 2013), (Wang, Chi, Zhou, & Wong, 2015), (Wei, Chow, & Chan, 2015) , (Kacem, N'cir, & Essoussi, 2015), (Ji et al., 2015), (Ren, Liu, Wang, & Pan, 2016), (Ahmad & Hashmi, 2016), (Wangchamhan, Chiewchanwattana, & Sunat, 2017), (Sangam & Om, 2018) , (Jang, Kim, & Jung, 2018), (Lakshmi, Shanthi, & Parvathavarthini, 2018)
2.	Hierarchical approach for mixed data clustering	(Chiu, Fang, Chen, Wang, & Jeris, 2001), (Cen Li & Biswas, 2002), (Chae, Kim, & Yang, 2006), (Hsu et al., 2007), (Hsu & Chen, 2007), (Hsu & Huang, 2008), (Shih, Jheng, & Lai, 2010), (Lim, Kim, & McLeod, 2012)
3.	Density based approach for mixed data clustering	(Chen & He, 2016), (Jinyin, Xiang, Haibing, & Xintong, 2017), (Liu, Huang, & Shen, 2017), (Liu et al., 2017), (Shekhawat & Sharma, 2017), (Duan, Gou, Yang, & Chen, 2019),

From Table 2.1, it is observed that most of the works in mixed-data clustering follow either a partition or hierarchical clustering approach. In the next section, presents a detailed discussion on these approaches.

2.3. An overview on the types of data

Real datasets often consist of mixed variables, that is, some variables are quantitative or numerical and others are qualitative or categorical. The numerical data type is measurable such as a person's height and salary. It further consists of discrete and numerical data type. A discrete data value can be counted but it cannot be measured. For instance, they take on possible finite values that can be listed such as $0,1,\dots,n$ or possible infinite values such as $0,1,\dots,\infty$. A numerical data value is a measurable quantity.

However, their possible values cannot be counted, rather can only be described using intervals on the real number line. The categorical data type is represented either as nominal or ordinal. The nominal data values represent discrete units and are used to label variables. They have no quantitative value nor any order to them. The ordinal data type represents discrete ordered units. They are nearly the same as nominal data type except that ordering matters for this data type (Ahmad & Khan, 2019).

2.4. An overview of mixed data unsupervised algorithms

In this section a comprehensive discussion on several unsupervised algorithms to process mixed data are presented. Their benefits and drawbacks are also highlighted.

2.4.1. Partition-based clustering

In partition-based clustering, the most famous algorithm for mixed data is the K-prototypes (Huang, 1997), which integrates the K-means and the K-modes algorithm. Feature selection is determined by an agglomerative similarity method for both numerical and categorical features between variables and cluster centres. For numerical data, the distance measured is the square Euclidean distance and for categorical data, that is the number of incorrect groups between the data points and the initial cluster centres. A weight w_i is added to the categorical variables so as to avoid bias in variable selection by the algorithm.

A partition clustering algorithm based on K-means for analysing mixed data called the K-means Clustering for Mixed Datasets (KMCMD) was suggested (Ahmad & Dey, 2007). This algorithm proposed a novel method and rate objective dependent on the recurrence of data features. The algorithm begins by randomly assigning a cluster to all objects. Then the cluster centre is calculated by assigning each object to the nearest cluster. The cluster centre is recalculated each time a new object is included. The assignment of objects and recalculation of cluster centres are repeated until the objects do

not modify their clusters. Unlike K-prototypes algorithm, the importance of feature is computed by discretizing the numeric features.

A weighted K-means clustering algorithm for mixed data was suggested (Modha & Spangler, 2003). This method works by arranging features in disparate space. They have suggested a measurement to calculate the occurrence of dissimilarity in two abstract features in a data environment. The dissimilarity in disparate feature set are collated to calculate mass of each feature. Their proposed algorithm distinguishes two feature spaces: namely the numerical feature space and the categorical feature space. Their proposed algorithm applies scaling to numeric features and applies a 1-in n-based representations in every n-categorical feature. A Euclidean distance that is squared is computed to measure the similarity within numerical features and the "cosine distance" is calculated to measure the similarity within categorical features. The drawback of their proposed method is a lack of comparative discussion with other unsupervised algorithms. Their method linearly scales the numeric feature by first removing the mean and then dividing the resultant with the standard deviation in every 1 in q^{th} representation for all q^{th} categorical feature. For measuring the distance, the squared Euclidean distance are computed for the numerical feature, and the cosine distance metric is calculated for the categorical feature. However, the study lacks in the aspect of no comparison with other similar clustering algorithms.

The researchers used the distance algorithm given by Ahmad and Dey (2007) to propose an unsupervised approach for streaming data contained mixed data features (Chen & He, 2016). They integrated micro-clusters in their proposed approach. A micro-cluster is used in data streams to efficiently compress the data. In this method, the midpoint of cluster is calculated to partition the data. This algorithm works by applying two different types of threshold, firstly the decay threshold followed by the density threshold. The density threshold observes the significance between the historical data

with reference to the existing cluster. If there is a significant difference, it then applies the density threshold to make a demarcation between the dense and the sparsely populated micro-cluster. The drawback with this method is the parameter-optimization.

Another group of researchers applied the algorithm given by Ahmad & Dey (2007) to suggest an unsupervised method to process mixed data (Ren et al., 2016). They applied the Euclidean distance between numerical variables and the Hamming distance for categorical variables. A kernel-based Gaussian to compute the distance within features was applied. Another group of researchers combined the theorem of cluster centrality proposed by Ahmad & Dey (2007) and a feature significance method proposed by Huang & Chow (2005) to develop and propose a novel rate objective function for mixed data clustering (Ji et al., 2013). The features were randomly selected for significance, leading to iteration-based update. However, the drawback with this approach is the randomness. Because of this randomness the algorithm is unable to reproduce results that are consistent with varied iteration cycles.

Recently, Sangam & Om (2018), suggested a K-prototype based distance for mixed data clustering. Their suggested approach assigns weights to categorical features that are dependent upon the feature frequency clustering disparity. The weighted Hamming distance was applied to measure the distance between categorical variables and the Minkowski distance was applied to the numeric variables. They have claimed that their suggested method is better than the K-prototype method (Huang, 1997). And have used accuracy, an external cluster validation metric for validating their results, which is the same as the Ji et al., (2013) paper. Besides this drawback, there is no discussion on the treatment of factor levels for a categorical variable. Moreover, they have not discussed the method to identify possible number of clusters to validate their approach. In the absence of not knowing the determinant number of groups in dataset, it's not possible to reproduce their results.

The researchers modified a "fast-genetic K-means" unsupervised learning method (FGKA) for mixed data (Roy & Sharma, 2010). Their proposed approach curtails the absolute variation between cluster centres by using the distance method (Ahmad & Dey, 2007). Although, the authors claim that their proposed algorithm outperforms the FGKA algorithm, but it lacks the explicability of the modification which is the central idea of their suggested approach in FGKA (that can only process numerical data).

A group of researchers suggested a partition dependent iterative unsupervised learning method for mixed data, that derives its inspiration from the unsupervised K-means clustering algorithm (Chiodi, 1990). A function applying a predetermined cost would penalize the occurrence of high-valued numerical features was applied in their proposed algorithm to balance the features. Chiodi (1990) also applied the Euclidean distance method to measure the similarity between numeric features and the Hamming distance was applied to measure the similarity between categorical features. The mean for numeric variable were used and the recurrent dependent spread was used for the categorical variables. The method was administered to an andriatric dataset. The parallelization of the K-prototypes algorithm by Huang (1997) was suggested (Kacem et al., 2015). This algorithm uses the MapReduce framework given by Dean & Ghemawat (2010) for parallelization. Recently, another group of researchers applied a method dependent on grid-based algorithm for the K-prototypes algorithm (Jang et al., 2018). For experimentation a geographically distributed dataset comprising of numerical and categorical data points were used. They have shown their results to be better than the K-prototype algorithm in terms of computational speed.

The other partition based mixed data clustering approaches involve data conversion. For instance, the researchers developed an algorithm which uses the concept of polar data points to convert variables that are categorical in nature into numerical format to be applied with K-means clustering to group the features (Barcelo-Rico & Diez,

2012). Wang et al. (2015) proposed an unsupervised learning method that was context-aware to mixed data. Their algorithm computed the relation of numerical and categorical features that was individually calculated. These determined the possibility of correlation within the numeric instantiation for mixed data. Then the partition clustering algorithm, the K-means was subtly injected to determine the similarity between the variables. Conceptually speaking, this is a smart approach which takes into account the relationship between the features. It is akin to measuring correlation and collinearity. And their experiments suggest their proposed algorithm showed good results for mixed data clustering. Wei et al (2015), suggested an approach that exploited the property of information present mutually among the variables for unsupervised feature selection for mixed data. Their approach is similar to that of Rico & Diez (2012) in which they too have converted the categorical data points into numeric-instantiations which were then subjected to the K-means algorithm.

So far, the hard partition-based data clustering is discussed. Fuzzy clustering represent the soft-data clustering approaches, wherein an feature can lie in more than one cluster and this affinity between its dependence is based on the membership degree within the cluster (Yang, 1993). In literature, there exist several fuzzy clustering methods for mixed data clustering. The researchers apply a dynamically computed probability based distance method such that it determine weights of numerical variables and distance amongst pairs of categorical variables (Ahmad & Dey, 2005). This fuzzy distance measure is collated within cluster's centre method given by researchers El-Sonbaty & Ismail (1998) to propose a fuzzy C-means (FCM) method for handling mixed data.

The K-Centers algorithm, proposed a novel distance measure to which considers different frequencies for feature value on cluster centres. This algorithm works in two ways. First, it initializes the cluster centre. Next, it will calculate the membership matrix. The algorithm updates the membership matrix and minimizes the rate objective to find a

new cluster centre. If an object belongs to only one cluster, then it is called Hard K-Centers clustering. If an object belongs to several clusters, then it is called Fuzzy K-Centers clustering. The cluster centre is computed repeatedly until rate objective cannot be minimized further. The drawback with this algorithm is that firstly, it cannot deal with outliers; secondly, it requires a user defined parameter for the initial number of clusters; and finally it cannot guarantee a local optimum solution (Zhao et al., 2007).

KL-FC-GM algorithm is an extended version of the K-prototypes algorithm (Chatzis, 2011). This algorithm is based on the fuzzy principle of uncertainty. It works by employing a probabilistic based dissimilarity distance measure. However, when the data dimension becomes large, the dissimilarity computation costs much more time. Zheng et al (2010) developed an algorithm called, Evolutionary K-Prototypes (EKP). It has a global search capability which was absent in the initial K-prototypes algorithm. Hsu & Chen (2007) proposed a variance and entropy-based clustering algorithm called CAVE for handling mixed data. However, the drawback pertaining to this method is that it determines a span hierarchy for each categorical feature, and this requires domain expertise.

A modified approach to the traditional K-prototypes algorithm was suggested by researchers (Ji et al., 2013). It worked by injecting a novel distance method for computing the difference between variables of different data types and their associated groups. Their suggested method exploits the fuzzy spread frequency of centrally located variables within a group. This approach is different from the traditional mode-based clustering aimed at finding similar categorical variables.

Another group of researchers modified the K-prototypes algorithm by discrete interval determination thereby removing the issues generated by conditional complementary entropy. A primary disadvantage of K-prototypes and K-modes algorithm

were that both will modify the cluster centres depending on the maximum frequency of feature values (Ji et al, 2015). This causes the cluster centres to ignore the significant value of other features subsequently degrading the cluster accuracy (Gu et al., 2018). The Partition Around Medoids (PAM) is a variant of the K-means algorithm that is applicable to a wide-range of applications like text mining, image analysis, bioinformatics etc. The complexity of the PAM method is O (k (n-k) ^{2}). This infers for large values of n and k, the computation complexity is very high (Kaufman & Rouseau, 2009). CLARA (Clustering for LArge Applications) was introduced to solve the inherent computational complexity drawbacks of the PAM algorithm. CLARA extends the PAM algorithm to cater for high-dimensional datasets. So, while it's efficient in processing large datasets, its efficiency drops if the dataset is biased (Kaufman & Rousseeuw, 2009).

2.4.1.1. Cluster Centre Initialization

A well-known issue with partition-based clustering algorithm is the problem of Cluster Centre Initialization (CCI). Generally, in partition clustering algorithms the initial number of clusters are selected at random. This causes differing clustering results with differing algorithm execution cycles, even though for the same dataset. Therefore, researchers find it difficult to reproduce the results. In literature, there exists several research papers, wherein the researchers have attempted to address this issue. Therefore, this sub-section gives a brief discussion of such approaches.

The researcher Ji et al (2015) suggested a method that can initialise cluster centres for the k-means clustering algorithm to work for mixed datasets. Their proposed idea was to determine median for numerical features by determining the spatial relationship between the data-points, and by exploiting the idea of adjacent feature occurring together. The median for numerical features and the distance between them was calculated to be the preliminary cluster focal points. The drawback of their algorithm is its computational

complexity that is quadratic in nature, as compared to the successive computational intricacies of the K-means-type clustering algorithm.

Several research works have been undertaken to initialize the cluster centres for the k-means algorithm (MacQueen, 1967). The researcher Forgy (1965) developed a CCI method in which the data points were initially randomly assigned to any of *k* clusters. Thereafter, the average of the data points of the clusters were taken as the primary cluster centres. These were then passed into the partition clustering algorithm as possible number of clusters. Another method suggested by Jancey (1966) that was rather crude in working, was to assign a machine defined cluster centre generated randomly from the data points within a given data space. In another study, MacQueen (1967), suggested two diverse approaches for CCI step; their primary method was to pick a random set of *k* data points from a dataset and assign them as the initial cluster centres. Their second approach was to randomly select a subset of *k* data objects and assign them as the initial cluster centres. The underlying assumption for this second approach was that randomness might pick initial good cluster centres. And this assumption although defective in nature has become the standard for the k-means clustering algorithm (Bradley and Fayyad, 1998). Its defective in nature, because it does not guarantee the reproducibility of results and in some other cases, it might even select the outliers as the initial cluster centres, which is again catastrophic. In another study proposed by Ball and Hall (BH) (Ball and Hall, 1967), the approach initially identifies the cluster centres and then picks a data point that is *T* distance far from the next cluster centre when compared to the initial cluster centre. It is an iterative process that continues and converges till the predefined number of *k* cluster centres is obtained. Unlike, the BH method there is another approach called the Simple Cluster Seeking (SCS) (Celebi, Kingravi, Vella, 2013) that picks up the first data object in the dataset as the CCI. But then again, this approach is plagued by the presence of outliers in the dataset. The Maxmin method (Gonzales, 1985) randomly selects a data

point from a given dataset as the initial cluster centre. It then compares this data point with others in the dataset by calculating the highest distance measurement between the existing cluster centroids and continues the iterations until k centres are obtained. The researcher Al-Daoud proposed two approaches, namely the Al-Daoud method 1 (AD1) (Al Daoud and Roberts, 1996) and the Al-Daoud method 2 (AD2) (Al Daoud, 2005). The AD1 approach uniformly segregates a given dataset into a pre-specified number of disjoint hyper-cubes, which are then randomly allocated to be the initial cluster centres. In contrast the other approach AD2, begins by determining the ranks for all data points in a dataset basis of their variance. The data point with the maximum variance is then selected and assigned to k groups along the same data point. Finally, the algorithm AD2 converges by applying the median of data points as the initial cluster centres. The kmeans++ approach (Arthur and Vassilvitskii, 2007) integrates the MacQueen's second method with the Maxmin method to initialise the cluster centres. Another algorithm called the Cluster Centre Initialization Algorithm (CCA) (Khan & Ahmad, 2004) begins by selecting the $k^i > k$ centres from the centroids obtained through the k-means algorithm (MacQueen, 1967) on each data point, to merge similar centres to formulate k initial cluster centres. The Redmond and Heneghan's method (Redmond & Heneghan, 2007) method applies the notion of *kd-tree* to compute the density. It then applies a modified Maxmin method for initialising the cluster centres. The researchers Cao *et al* (2009) postulated an initialization approach which derived its inspiration from the neighbourhood-dependent rough-set model. Another group of researchers, Yi *et al* (2010) suggested that data objects which belonged to high density areas in a cluster should be chosen as the initial cluster centres. Kumar, Chabbra and Kumar (2011) suggested to apply a biography-based initialization approach for determining the cluster centre initialization.

Similarly, the literature is replete with abundant methods to perform the cluster centre initialization for the k-modes algorithm. Huang suggested two approaches (Huang, 1998): the primary approach is a brute-force method that takes the first *k* distinct objects as the initial cluster centres; the second approach is begins by initialising the frequent categories to *k* distinct data points as the initial cluster centres. This approach is quite similar to approach of the Al-Daoud approach and the Maxmin method. Sun *et al.* (2002), suggested a cluster initialisation approach which utilised an incremental refinement process that was formulated by researchers Bradley and Fayyad (1998) to improve the meaningfulness of the cluster centres (Khan & Ahmad, 2013). The researchers Khan and Ahmad (2003) combined the two-distance metrics namely the Hamming distance with data compression method (Mitra, Murthy & Pal, 2002) to compute the initial cluster centres. This approach works by calculating a random data point as the initial cluster centre, and then designates a data point at the farthest distance to the nearest cluster and appoints it as the next cluster centre. It is an iterative process and continues until the expected *k* clusters are obtained. The second approach, calculates a scoring method to gauge the data points with the highest score and picks it up as the initial cluster centre. Wu *et al.* (2007), applied the notion of density to obtain the cluster centres. The drawback of this method is the process of random sampling the dataset, which causes unstable and non-reproducible clustering results. Cao *et al.,* (2009) calculated the distance between the data points and then ingrained it to the density of the data points. The resultant was then suggested as the initial cluster centres. Their proposed approach initially assigns a boundary between the data points on basis of their variance. Khan and Kant (2007) suggested the usage of evidence accumulation theory (Khan and Ahmad, 2013) to compute the CCI. In this proposed approach, the k-modes algorithm was randomly initialised and executed *n* times to yield a mode-pool. From this pool, distinct modes were selected and designated as the initial cluster centres. Furthermore, the researchers Khan

and Ahmad (2013), suggested the application of “multiple-attribute clustering”, as the method for cluster centre initialisation.

The researcher Forgy (1965) suggested that every data point present in a cluster must have a uniform distribution. Also the data points in such groups will then be designated randomly to the cluster. The focal point of the groups are determined by their centroids. Later the researcher Anderberg (1973) asserted that affixing randomness to cluster distribution has no internal consistency to a dataset. The researcher Jancey (1966) suggested to designate a user defined value to each cluster. Again, the researcher Anderberg (1973) dispelled this idea by suggesting that this approach was not appropriate to use because it would cause unequal assignment of data points to clusters, and it can further aggravate the problem by the initialisation of barren clusters. The researcher MacQueen (1967) suggested that cluster members must be assigned membership basis of their data location within the dataset space. MacQueen suggested that this approach was better than the random data point assignment to a cluster center because the data points will not be randomly assigned. However, the researcher Anderberg (1973) was quick to point out that this approach was faulty because it had the potential to assign outliers as cluster centers. Later the researchers Bradley & Fayyad (1998), suggested an improvement to MacQueen's approach. They suggested the algorithm to be executed several times and the average of the result be taken as the cluster centre. The method given by researchers Tou & Gonzales (1974), is very similar to the Ball and Hall's method with a minor difference, that the initial cluster center s affixed the first data point in a given dataset. The researcher Spath (1977) suggested a similar method to Forgy's but with a minor difference, the data points are to be cyclically ascribed to a cluster to avoid the problem of sequential data point placement present in the Forgy's method. The researchers Gonzales (1985) and Katsavounidis et al. (1994) suggested to apply the maxmin method to a group of clusters such that, the cluster center is assigned the highest

minimal distance. This method is motivated by the highest Euclidean distance to be computed as the initial cluster focal point.

The researcher Al-Daoud & Roberts (1996) density-based method has been criticised in literature for its two fundamental issues. The first issue is that it's difficult to determine an appropriate number of data points present in a data space, and, is the storage complexity of the algorithm which is huge. Besides this, another major drawback of their approach is data point sorting method. Essentially, their method is biased for a multidimensional dataset. This infers to the sorting approach their method undertakes. The way it works is by initially all data points are sorted based on their frequency count for categorical data points and variance is considered for the numerical data points. This sorting occurs only in one-dimension. The sorting approach disregards the other dimensions if any present in the dataset.

Pizzuti, Talia, and Vonella (1999) proposed a major improvement to Al-Daoud's density-based approach by suggesting the application of a grid approach. Their idea was to split the data space into a predefined number of disjoint spaces. Then, chose a representative data point from a densely populated data sub-space.

A dimensionality reduction method based around the famous Principal Component Analysis was suggested by the researchers Su & Dy (2007). Although its PCA based but it applies a hierarchical approach for reducing the data dimensionality. Their idea was to initially collate all data points into a subspace and then calculate it sum of square errors. This is repeated for each cluster. Then the clusters are split into sub groups based on their orthogonal data distribution within the clusters such that cluster centers align with the primary Eigen components of the dataset. Furthermore, this process continues until a predesignated set of groups are found. The researchers had also

suggested another distance calculation method which was dependent on the similarly connected data points.

A group of researchers (Lu, Tang, Tang, & Yang, 2008) applied a two-step hierarchical approach for cluster center initialization. Initially the categorical data points were hard encoded into number format. These were then assigned at the initial level zero of the data hierarchy. Then the data points were simultaneously computed on median until an initial batch of data points were obtained. Then the K-means algorithm was applied to partition the data into sub-spaces. The potential drawback with this method was that it could not handle high dimensional datasets (Lu et al., 2008). Onoda et al.'s method (Onoda, Sakai, & Yamada, 2012) proposed a novel approach for cluster center initialization. Their idea was to initially compute the n independent components from the data point space. Then identify the data points which had the minimum cosine distance from the n independent components and designate them as the cluster centers. The researcher Hartigan & Wong (1979) suggested to initially sort the dataset dependent on their intra distances between each other. This method improved the MacQueen's first approach which was data distribution plays a pivotal role in producing cohesive clusters. But this method had a huge time complexity which was attributed to the high number of sorting operations involved in it. Moreover, it was unclear from this approach the type of data sorting method involved. For instance, was the data sorting quicksort based or merge sort or random sort.

The researchers Redmond & Heneghan (2007) proposed a k-dimensional tree method for the cluster center initialisation problem. The idea was to first arrange all data points in a data space in increasing order of their individual densities into a k-dimensional tree. Then they applied a customised maximum-minimum method to determine n-data centers from the leaves of the tree. However, the computational cost of this algorithm was high and same as the MacQueen's first approach. A group of researchers improved the

"Local Outlier Factor (LOF)" approach proposed by (Breunig, Kriegel, Ng, & Sander, 2000) by removing the outliers that were selected as the initial cluster centers. Their proposed idea was as follows: first assign an initial cluster center. Then sort all data points that fall within the cluster center range on basis of their distance from the cluster center. Arrange all such data points in a decreasing format. Finally, the method will go through all the sorted data points to choose the data point whose LOF value is less than or equal to 1 as the new cluster center. Again, the computational cost of this method is heavily dependent on the sorting and the data dimensionality.

The researcher Astrahan's approach (Astrahan, 1970) applies two disparate distance measures d1 and d2. Initially, the number of data points within a given distance d1 are collated and their individual data density is computed. Next, the data points are sorted and arranged on the basis of their decreasing density range. Then the data point which is on the top of the line with the highest density is assigned to be the initial cluster center. This process continues again, by sorting the remaining data points and arranging them on basis of their decreasing density range. Again, a new cluster center is chosen from the remainder of these data points with reference to the earlier cluster center. Furthermore, in this algorithm if there happens to be more than a predefined set of clusters, then the hierarchical grouping method is applied such that the number of clusters remain equal to the predefined number of clusters. The fundamental problem with this method is its high sensitivity to the initial distance measures d1 and d2. A group of researchers in 2009, improved the Astrahan's density-based method for cluster center initialization by incorporating it into a rough-set algorithm which was based on the premise of neighbourhood location. The idea was a group of data points in a data space were perceived as a group g. The d-data point in the adjacent group or the neighbourhood was about m-distance away from the data point in g. Then the central location of m with respect to g was defined as cohesion and the distance between m and g was defined as

coupling. Their proposed approach suggests to initially sort the data points in the decreasing range of their cohesion. Then assign the data point with the highest cohesion as the initial cluster center. The approach, then iterates over the sorted data points and chooses the data point which has a coupling lesser than the initial designated cluster center. Once again, the drawback of this approach is its computational complexity which is higher if the number of data points in the neighbourhood are high in number. Lance and Williams (1967) proposed a approach to quantify the initial number of groups for the k-means method. They suggested the data points in a given dataset first be subjected to a hierarchical clustering algorithm. The resultant number of groups can then be considered as the cluster centers for the k-means. Although, this approach had a high quadratic computational complexity because of applying the hierarchical clustering algorithm, but it's been widely recommended (Milligan, 1980). The Kaufman and Rousseeuw's (1990) method was dependent on reducing the sum of squared distance. This approach too has a quadratic complexity because pair-based distance is computed between data points for all algorithmic iterations. The researchers Linde, Buzo, & Gray (1980) suggested a binary-splitting approach. The idea was to initially select n random data points from a given dataset. Then traverser through each of the data points such that the distance between a given pair of data points is computed and split into binary clusters by applying the k-means algorithm. This approach suffers from two problems: primarily, the initial split criterion which coerces a given data point into binary format is random in nature, as suggested by Huang & Harris (1993). And the second issue is, the approach is computationally expensive. Another method suggested by Huang & Harris (1993) was the binary demarcation of a directed cyclic search algorithm. They suggested it to be an improvement over their earlier approach that split the data point's basis of their binary properties. This approach used the PCA method to determine relevant data points in a given data space as well as for dimensionality reduction. But because of the usage of

Eigen vectors in the calculation of PCA, the algorithmic complexity of this approach is huge.

There exists some annealing based simulated algorithms in literature for cluster centre initialisation. Such methods operate by initially setting a randomly selected base population of data points, which are then grouped by applying the k-means partition method. The process is repeated for k times where k is a predefined number. There are primarily two major drawbacks with these methods" primarily, the number of tuning parameters is very high (Jain et al., 1999), such that tuning the parameters to attain a substantial degree of confidence is computationally expensive. The second issue is because of the high number of tuning parameters, manifold iterations occur which memory is consuming even for a small dimensional dataset. But with rapid improvement in algorithms, the researchers have developed methods that substantially curtail the algorithmic complexity of annealing methods by minimizing the sum of square errors for low dimensional dataset (Aloise, Hansen, & Liberti, 2010).

2.4.1.2. Number of Clusters

Perhaps, one of the fundamental drawbacks in using a partition-based clustering algorithm is not knowing how many groups exist in a dataset. To determine the number of groups in a dataset, invigorates the employability of a user-defined number or the usage of algorithms like the Elbow method (Kaufman & Rousseeuw, 2009). However, approaches such as these often are unable to provide a decisive number of groups that exactly represent the distribution of data points in a dataset.

The unsupervised learning methods that are partition-based for either numeric or categorical data like K-means and K-modes suffers from many impediments notably determining an appropriate number of clusters. This anomaly is inherited by the mixed data too. By applying the concept of density peaks two diverse group of researchers

suggested an approach for calculating the preliminary cluster focal point for mixed data (Rodriguez & Laio, 2014),(Jinyin et al., 2017). Their argument was clusters with elevated peaks were indicators of possible cluster centres. Again, the drawback of their algorithm is its computational complexity that is quadratic in nature. Wangchamhan et al (2017) applied a "league championship" search method which had the K-means algorithm integrated in it and was used to determine the inceptive group focal points. For distance measure, the Gower (1971) distance was used. The problem with their proposed approach was parameter selection. Lakshmi et al (2018) applied the crow-tuning algorithm to determine the inceptive group focal points for the K-prototypes method. The salient feature of this algorithm is that its performance is better than the K-prototypes method that works on the principle of random cluster focal points. But it suffers from the parameter selection step which is a crucial step for the crow-tuning algorithm, and thus it leads to the problem of the same clustering result that are irreproducible if disparate parameters are applied. Ahmad & Hashmi (2016) combined the distance measure and cluster definition proposed by Ahmad & Dey (2007), with a function that assigns a predetermined cost to K-harmonic clustering proposed by Zhang (2001), to enhance the K-harmonic grouping for data which is mixed in nature. Experimental evidence suggests the proposed approach was optimal for determining the clustering indexes in contrast to the K-means method for mixed datasets. A group of researchers collated two separate algorithms together. The first algorithm was the unsupervised K-prototypes and the other was an evolutionary algorithm (EA). They exploited the searching mechanism of the EA algorithm to improve upon the sensitivity of clusters. This approach helped them extract a good cluster performance. A distance measure suggested by Rahman & Islam (2012) was applied to measure the distance for categorical features. Although, the algorithm has shown good results, but it has a quadratic computational complexity to it (Zheng et al., 2010).

For the partition clustering algorithms to work, the number of clusters need to be predesignated. The numeric value for the possible number of groups or clusters is often derived by executing similar algorithms or is defined randomly. Although, the efficacy of such approaches to determine the number of groups rules out the chance of cluster purity. And thus, this problem percolates to partition-based algorithms for mixed dataset too. Liang et al (2012) suggested the application of an index derived from validating the test cases, that was used to determine the number of clusters for mixed data. Their proposed algorithm consists of two interlinked modules: the primary module analysed the numeric or continuous variable and the secondary module analysed the categorical variables. The Gluck and Corter (1985) method is applied to the categorical features and a utility method suggested by Mirkin (2001) was applied for processing the numeric features. Weights are assigned based on the occurrence frequency of numeric variables or categorical variables. Subsequently, this approach was repeated for a varied number of random clusters. The clusters number which maximize the cluster validity index were designated as the optimum index of clusters. To evaluate the cluster validity, the algorithm uses the Renyi entropy (1961) to process the numerical features and the inverse entropy given by Liang et al (2002) for the categorical features. Thereafter, the K-prototypes algorithm is applied for clustering. Although, the proposed approach was efficient in predetermining the cluster index but its efficacy is questionable as it was tested on datasets for which the number of groups was known in advance (Liang et al., 2012).

The researchers Milligan & Cooper (1985) discussed a detailed overview of thirty different types of internal clustering indices that can be applied to determine the number of clusters in a given dataset. It was suggested that the Calinski-Harabasz index was superior to others. Salvador and Chan (2005) suggested the "L" approach to determine the "knee-point" in a clustering graph. The idea was the L algorithm is formulated by a pair of lines that evolve from either side of an evaluator graph, which is a close fit to a

curve. The merging point of these lines is designated the "knee". Their experiments have shown that the "L" method was better than the gap statistic method, which was only able to determine one correct number of groups out of the seven experiments conducted by them.

The researchers Pedersen & Kulkarni (2006), suggested the usage of a software application that could help compute the possible number of groups in a given dataset. They actually applied four unsupervised methods given by Hartigan & Wong (1979), Mojena (1977), Dice coefficient (1945) and an enhanced gap statistic measure. The researchers found that off these four different unsupervised methods, the enhanced gap statistic measure yielded the best results. Their justification to better results obtained by the enhanced gap statistic measure was that it did not waste computation power in determining the elbow point which the other three approaches were heavily dependent upon.

A group of researchers (Charrad et al., 2012) developed a package using the R programming language called the “NbClust”. The software package consisted of 30 validity indexes. Essentially the package computes the elbow method by applying a wide range of methods classical such as the knee or elbow finding approach to the modern approaches like the maximum indices value, the minimal indices value, and the maximal inequality indicator, the maximum difference between hierarchical levels, the graphical method and several others. The package works by determining the number of groups in a given dataset by applying all 30 cluster validity indices. Then, the resultant is chosen based on the maximal occurrence frequency of a given number of groups found by the 30 cluster validation indices. This repeated computation forces the algorithm to have a high complexity. The researchers Zhang et al. (2014), suggested a weight network based fuzzy clustering validation index that could compute the initial number of groups in the dataset. The researchers too applied various grouping methods like the K-means, hierarchical, EM

and fuzzy C means methods. The researchers validated their results by applying the DBI index on nine artificial and related real-world datasets.

2.4.2. Hierarchical-based clustering

The hierarchical clustering methods develop ranking based groups organized in an increasing or decreasing order. For clusters to form, the hierarchical algorithm must fulfil the following conditions:

i. Affinity model - is developed by determining the likeness across pairs of mixed data objects, where the choice of affinity model determines the outline of a cluster.

ii. Linkage criterion- This calculates the distance between observation pairs by determining the pairwise distance between the observations.

The singular drawback with hierarchical clustering algorithms is the huge time convolution of $O(n^3)$ and consumes heavy memory usage On^2, as n constitutes possible counted variables. A review of hierarchical clustering approaches for processing mixed data is presented next.

The researchers (Philip & Ottaway, 1983) applied the Gower's similarity coefficient (Gower, 1971) for calculating the feature similarity by bifurcating the features into categorical and numerical data types. To compute the likeness across categorical data points, they applied the Hamming distance and to compute the likeness across the numerical data points, they applied a custom function. This function was designed to work in such a way that similarity between same features was assigned with the value of 1, whereas the difference was assigned the highest difference (is the dissimilarity between the highest and the lowest value in a variable), the similarity was assigned 0. Furthermore, this custom function calculated the sum of similarity for all numerical variable as the relatedness between two variable values for a numeric data environment. Regarding the

categorical data points, the similarity between variables was calculated by assigning a predetermined function value. Finally, the similarity amongst the categorical features and the similarity amongst the numerical feature space was added together to develop the likeness across two mixed data points. Thereafter, the hierarchical agglomerative clustering was applied for developing the groups (Philip & Ottaway, 1983).

A group of researchers created a custom similarity method to calculate the likeness across two disparate groups of mixed data. The log-likelihood distance method was applied as a distance function for coalescing the two groups of mixed data. Essentially, the BIRCH clustering algorithm (Zhang, Ramakrishnan, & Livny, 1996) was collated with their proposed distance measure to create a method for computing the likeness across mixed data points (Chiu et al., 2001). However, another group of researchers proposed an idea which had its roots in the theory of conceptual hierarchy. This theory suggested the idea of opinion nodes and links (Han & Fu, 1994) and (Han, Cai, & Cercone, 1993). Their proposed idea was the higher-level nodes contain the general concepts whereas the lower-level nodes contain the specific concepts. So, the categorical features are to be represented in a tree structure, such that the leaves denote the presence of a categorical data point. Thereafter, the distance between any two given feature values is computed by applying the hierarchical correlated distance between them. Finally, an aggregated hierarchical clustering algorithm that was proposed (Hsu et al., 2007) and was applied to a matrix of distance calculated earlier to obtain the clusters. The drawback in their proposed approach is that it was dependent on domain knowledge to develop the distance hierarchies for the categorical features (Hsu et al., 2007). A novel similarity measure for mixed data clustering. This measure applied the concept of distribution for numerical features that decays with distance based ranking for the categorical features. Thereafter, they applied cumulative grouping to group the mixed data points (Liang et al., 2012). The concept of adaptive resonance theory (ART) to group

mixed feature values by calculating the hierarchical span between the mixed data points as input to the network. They suggested that their experimental results provided better results when compared to the partition clustering algorithm the K-prototypes (Hsu & Huang, 2008). The researchers proposed an algorithm in which they converted the categorical variables in a mixed dataset to numeric variables on basis of their co-occurrence frequencies (Shih et al., 2010). Thereafter, the numeric features and the converted categorical to numeric features were grouped using the hierarchical clustering algorithm proposed (Hsu & Huang, 2008). The drawback of their proposed approach was data conversion incurs data loss and this was not discussed in their paper (Shih et al., 2010).

Another group of researchers proposed an algorithm that began by partitioning the feature space into categorical features and numerical features. The partitioned feature space was then grouped separately to form clusters. To increase the associativity of the grouped data points, the grouped results were assimilated by infusing a predetermined distance metric to obtain a similarity matrix. Gower's similarity measure (Gower, 1971) was applied to ascribe balanced load to both categorical and numerical data points to mixed data. Thereafter, agglomerative clustering was applied to obtain the final clusters. The drawback in their proposed approach was that similarity matrix may be influenced by a particular data type. It was unclear from their proposed approach that how it would work for imbalanced mixed datasets (Lim et al., 2012). Table 2.2 provides a synopsis of the hierarchical clustering methods for mixed data.

Table 2.2: Summary of hierarchical clustering methods for mixed data

Author, year	Similarity measure	Clustering algorithm
Philip & Ottaway, 1983	Gower's similarity matrix (Gower, 1971)	Agglomerative hierarchical grouping approach
Chiu et al., 2001	Probabilistic model by applying a log-likelihood distance measure	BIRCH algorithm (Zhang et al., 1996)

Li & Biswas, 2002	Goodall similarity measure (Goodall, 1966)	Agglomerative hierarchical grouping with group-averagemeasure
Hsu et al., 2007	Distance hierarchy by applying concept hierarchy (Han & Fu, 1994),(Han et al., 1993)	Agglomerative hierarchical grouping approach
Hsu & Chen, 2007	Variance for numerical data points and entropy for the categorical data points (Hsu et al., 2007)	Incremental grouping approach
Hsu & Huang, 2008	Similarity measure proposed by Hsu and Chen (Hsu & Chen, 2007)	Adaptive resonance theory network (Carpenter & Grossberg, 2010)
Shih et al., 2010	Converted categorical data points into numerical data points	Hierarchical agglomerative grouping approach (Jain & Dubes, 1988)
Lim et al., 2012	Separate similarity measures for categorical and numerical data points space	Agglomerative hierarchical grouping approach
Chae et al., 2006	Modified Gower's similarity approach	Agglomerative hierarchical grouping approach

An unsupervised approach was proposed by researchers, that was dependent on the spread of data and its degeneration (Hsu & Chen, 2007). They termed their approach as CAVE. Their approach applies variance as a similarity measure between numeric variables and entropy for categorical variables. The drawback with this algorithm is that it builds a distance hierarchy for categorical variables which requires domain expertise. The Similarity-Based Agglomerative Clustering (SBAC) algorithm is a hierarchical clustering method (Li & Biswas, 2002). It applies the Goodall (Goodall, 1966) similarity measure to process numeric and categorical features. Thereafter, an agglomerative approach is used to build a tree-based structure. The SBAC is built upon the Unweighted Pair Group Method with Arithmetic (UPGMA) average (Goodall, 1966). The UPGMA algorithm applies a distance matrix pair for the collection of data objects. The distance among a couple of data objects is the counterpart to their measures of similarity values. At any given time, the lowest pairwise dissimilarity data objects of clusters are combined into a distinct group. The distance between the new cluster and the old clusters are defined

as the average distance between them. The computation of the dissimilarity measure is repeated until all the objects are combined in a single cluster. The termination of the cluster process outcomes in a dendrogram (or tree) where the leaf vertices will specify different data objects and root vertices specifies a group which contains the entire object.

2.4.3. Model-based clustering

The concept of model-based clustering is dependent on the statistical data distribution, which assumes that data point matches a model (Melnykov & Maitra, 2010). Since the models are user-defined therefore they incur the problem of yielding undesirable results, if inappropriate distance measure or any other parameter is incorrectly chosen by the user. Model-based clustering algorithms are typically feeble in performance when compared with the partition clustering algorithms (Melnykov & Maitra, 2010). A brief overview of major model-based clustering methods for mixed data is presented below.

The AUTOCLASS algorithm collates the finite distribution and Bayesian methods that are dependent in knowing the prior data distribution of individual feature (Cheeseman et al., 1988). This algorithm can group both categorical and numerical data type. Everitt (1988) proposed a clustering algorithm that applied model-based clustering to group both categorical and numerical data type occurring in a mixed data. The proposed algorithm would work only for binary or ordinal categorical feature. In this algorithm, the stable model is advanced to process mixed data by calculating starting indexes for the categorical variable set. The drawback of this algorithm is its increased computational cost. Another drawback of this algorithm is it only considers binary type ordinal feature set. It will not work for an ordinal feature set consisting of more than two categories. Moreover, this algorithm will not work for nominal categorical feature set. Thus, because of these aforementioned shortcomings, this algorithm can only be used for mixed data

with restricted number of variable data types. To resolve this issue, researchers extended the equivalent dependable Gaussian representation of extent sized fusion to compute the highest probability approximation for the selected variables in a feature set (Lawrence & Krzanowski, 1996). They assert to have obtained promising results and that their results have overcome the shortcomings of the AUTOCLASS method (Cheeseman et al., 1988). The authors have also suggested that their proposed algorithm can work for an arbitrary number of features present in mixed data. A quiescent class fusion approach for mixed data clustering was suggested (Moustaki & Papageorgiou, 2005). Their proposed algorithm converts the categorical features into binary features by a 1-in-n representative model. In categorical variables, a multinomial distribution is applied and a normal distribution to the numeric variables. The variables are deemed independent. However, their paper shows no evidence of statistically computing the independence of the features and this is a major shortcoming of their proposed approach.

An algorithm using latent variable model to group mixed data called CLUSMD was proposed (McParland & Gormley, 2016). The central premise for this algorithm was that Gaussian distributions when collated with latent variables for mixed data, then the results are better. The Expectation-Maximization (EM) algorithm was applied to determine the high variance features for the data. To deal with the categorical features, the Monte Carlo EM algorithm (McLachlan & Krishnan, 2007) was used. The authors claimed to have obtained better results using this approach, however, the drawback of their approach was that it becomes computationally expensive with an increase in feature number. To solve this problem suggested an unsupervised method to address mixed data with several variables by applying a Bayesian finite mixture model method (McParland & Gomley, 2016). This method applies the Gibbs sampling algorithm for estimating the number of relevant features to determine the optimum clustering model, an approximate Bayesian Information Markov Chain Criterion was applied. By using this conglomerate

approach, the authors claim to have obtained improved results for mixed dataset. Following their approach, a group of researchers, suggested to project the categorical features amongst the numerical feature subspace. By applying the PCA approach they demonstrated they were able to obtain promising results for mixed data clustering (Saâdaoui et al., 2015). A clustering algorithm dependent on Gaussian mixture copulas for mixed data clustering was developed (Rajan & Bhattacharya, 2016). A copula is termed a, "functions that join or couple multivariate distribution functions to their one-dimensional marginal distribution functions" and "distribution functions whose one-dimensional margins are uniform". Using the copula approach for mixed datasets they were able to model the feature dependencies to eliminate highly correlated features. Their method was shown to be better than the other model- based clustering algorithms. A group of researchers applied the concept of vines-copulas for mixed datasets (Tekumalla, Rajan, & Bhattacharyya, 2017). A vine copula, caters for a supple approach of applying a couple-wise inclination by using the stepwise collections of two variable binding, either of which can be associated to any bonded cohort thereby coalescing a wider variety of dependencies (Tekumalla et al., 2017). The core of their proposal applied the Dirichlet process mixture approach (Tekumalla et al., 2017). A fusion approach of Gaussian copulas for mixed data clustering was developed (Marbac, Biernacki, & Vandewalle, 2017). Their proposed approach integrates a Gaussian copula mixture by determining the correlation coefficient for feature pairs. Model selection is done by applying two information criteria: the Bayesian information criterion (Schwarz & Glassner, 2003) and integrated completed likelihood criterion (Biernacki, Celeux, & Govaert, 2000). The BIC is computed by applying a Metropolis-within-Gibbs sampler. Another group of researchers developed a semi-parabolic approach, "Kay-means for MIxed Large" (KAMILA) for grouping mixed data (Foss, Markatou, & Ray, 2018). This algorithm aimed to create a balanced proportion of numerical and variables that are categorical in

nature by integrating the K-means algorithm (for numeric data) and the Gaussian multinomial mixture models (Hunt & Jorgensen, 2011). Yet another group of researchers, integrated a mixture model into a fuzzy clustering algorithm for measuring the similarity between mixed data. This method also helped them to determine the cluster centres. Their idea was to apply probability inverse to data points occurring in a cluster by formulating the spatial relationship within the cluster centres and the feature points (Doring, Borgelt, & Kruse, 2004). The researcher developed a FCM based clustering algorithm for mixed data (Chatzis, 2011). This algorithm works by applying a probabilistic based dissimilarity function to a FCM based rate objective (Honda & Ichihashi, 2005). The unsupervised fuzzy approach was collated into a unified structure for handling mixed data (Pathak & Pal, 2016). Their idea was to segregate the numeric and categorical features apart. Then apply fuzzy clustering to the numeric features and mixture models to cluster categorical features (Bishop, 2006), (Chatzis, 2011). And to determine the possible number of cluster's they applied collaborative clustering approach (Pedrycz, 2002). Böhm et al. (2010) formulated an approach termed INTEGRATE to assimilate the information contained in mixed data. Their method used the concept of probability distributions from information theory to compose an idea that contained both numerical and categorical variable types and reduce the cost methods assigned to the concept of reduced description radius.

2.4.4. Neural network-based clustering

A considerable number of the research works related to neural network-based clustering applying mixed data is dependent on the application of Self Organizing Map (SOM) (Kohonen, 1982) & Adaptive Resonance Theory (ART) (Lam, Wei, & Wunsch, 2015). A SOM is a neural network (Kohonen, 1982). It functions by applying a non-linear scoping of the dataset to a lower-dimensional data point sub-space such that the unsupervised experiment can be conducted on this truncated feature space. The

inspiration of ART is derived from the functionality of the human brain. A potential drawback with SMO is its proclivity to create fixed-size maps. To overcome this drawback, an alternative approach in which the SMO growth was proposed (Alahakoon, Halgamuge, & Srinivasan, 2000). The SMO growth initialises itself with a small set of data points around which it develops an initial map. It works by automatic classification and prediction of objects dynamically (Grossberg, 2013). The central idea of ART's predictive approach is its proclivity to conduct a quick, gradual and stable increase, regardless of supervised or unsupervised analysis, when exposed to a dynamic environment considering that both SOM and ART methods can analyse and process numeric features only. They are unable to handle categorical features. To process categorical features using the SOM approach, researchers have mostly resorted to converting categorical features into numeric binary features (Prasad & Punithavalli, 2012), (Lam et al., 2015). And this is a major drawback with these algorithms. They do not consider the aspect of data loss in such transformations. A visualization-imbued SOM that stores the data structure and is better in performance than SOM was proposed (Yin, 2002). This approach was extended to group visualization-imbued SOM with a generalized SOM model to calculate the similarity for categorical features by applying the principle of generalization hierarchy was proposed. It comprises of knots or nodules and weighted links: the arbitrary generalizations are coded as top-level knots whereas specific generalisation is coded as knots to cluster mixed data. Experiments indicated to have obtained improved results (Hsu & Lin, 2006). The spatial-relation based grading are implemented to calculate the similarity amongst variables in the variable space (Hsu, 2006). In another experiment, the conceptual Self Organizing Map (SMO) was used to create an unsupervised approach for mixed data (Tai & Hsu, 2012). Some other researchers suggested a clustering algorithm by applying SMO approach, such that the Hamming distance was applied to categorical data points and the Euclidean measure was

applied to the numeric data points. However, the potential drawback with this approach was that it assigned higher weights to categorical features (Chen & Marques, 2005). For resolving this potential issue, a group of researchers modified the Hamming distance method such that categorical features were assigned equal weights (Coso et al., 2015). They claimed to have obtained improved findings as compared with the approach suggested (Chen & Marques, 2005).

The researchers applied an additive grouping method for mixed data by integrating the notion of self-adjusting and self-regulating neural network algorithm was developed (Noorbehbahani, Mousavi, & Mirzaei, 2015). They proposed a novel distance method for categorical features. Their idea was to calculate the interval range for categorical features based on their frequency of occurrence. This co-occurrence of feature values has already been discussed (Ahmad & Dey, 2007). But the potential drawback is the classification accuracy metric is not taken into account. Lam et al (2015) applied a clustering approach to yield a thinly dispersed for mixed data. A fuzzy adaptive resonance theory (ART) was applied to develop new features in their proposed approach. This approach begins by cluster centers of the data points, by randomly mixing and encoding them. They are then aligned to the original data points in the nascent data subspace. Thereafter, K-means approach was applied to group the elements in the new data point subspace. The researchers, Hsu & Huang (2008) applied the ART approach to develop a similarity-based distance-matrix for clustering data points by applying hierarchical clustering algorithm.

Besides the aforementioned clustering algorithms, there are other types of clustering algorithms in mixed data analysis that are inconspicuous and might not be well known. These are briefly presented. The spectral clustering methods execute dimensionality reduction by applying a similarity matrix of Eigen values (Ng, Jordan, & Weiss, 2002). Reducing the data in such a way, helps in determine cluster purity. In this

process, initially a similarity based distance matrix is calculated, and then a spectral grouping method is integrated to obtain the clusters. Luo, Kong & Li, (2006) developed a technique for computing similarity by applying an ensemble-based clustering method. Using this method, the likeness across two disparate data points is calculated independently for numeric and categorical data points. Thereafter, the two similarity matrices are combined to yield a common similarity index for a given set of features. And, finally the unsupervised spectral grouping is applied to this similarity matrix for acquiring cohesive groups. (David & Averbuch, 2012), applied a categorical spectral clustering approach to cluster mixed data. Their proposed approach works by converting the numerical data type variables into categorical features. It then uses the Calinski and Harabasz index for ascertaining the possible number of groups (Caliński & Harabasz, 1974). Subsequently, the unsupervised spectrum-based method is furnished to the converted variable data types. The researchers Niu et al (2015) proposed a grouping approach for handling mixed data. Their method aims to calculate similar matrix separately for both numeric and categorical data types. The concept of coupling relationship is applied to calculate the feature similarity. Both distance-matrixes were added by assigning an aggregated weight which was then used to compute the affinity grid for mixed data. This method was subjected to an e-learning dataset. They claimed to have obtained improved performance when compared with the K-prototypes method and the SpectralCAT approach (David & Averbuch, 2012). The idea behind subspace clustering is to identify groups in disparate subspaces in a group of data points. Ahmad & Dey (2011) applied a distance-based cost-function method. Using K-means as the clustering algorithm for subspace clustering, they applied it to a mixed data. Jia & Cheung (2017) applied an aggregated weighted grouping model to data points that utilized soft-subspace similarity matrix for mixed data. In their method, a weighted schema for numeric and categorical features is applied for measuring the contribution features. Plant

& Böhm (2011) created a clustering method called Interpretable Clustering of Numeric and Categorical Objects (ICONCO) that yields clusters of mixed data that can easily be interpreted. The algorithm works by applying data compression as in the minimal description length principle (Rissanen, 1978). The ICONCO method applies the requisite feature dependencies by utilizing the concept of linear modelling and subspace clustering for mixed data. The drawback of this algorithm is that it emphasises for a uniform distribution of categorical features with respect to the numeric features in a mixed data. The density-based clustering methods define clusters based on high density spaces. Du & Xue (2017) and Du, Xu, & Xue (2017) proposed an alternative distance measure for mixed data clustering. Their idea was to assign weights to categorical features and then collate it with the density-based clustering to determine the cluster numbers. The drawback of this approach is the selection of various parameters for algorithm tuning. Another clustering algorithm is conceptual clustering (Fisher, 1987). This type of algorithm develops a description of concepts for every generated cluster. It typically creates hierarchical category structures. The algorithm COBWEB applies a utility called as category utility measure (Gluck, 1985) to determine the relationship between clusters (Fisher, 1987). This utility measure can handle both numeric and categorical features presented in a mixed data. The algorithm COBWEB collates the COBWEB method and CLASSIT method to process numeric data points in the category utility algorithm. The drawback with this approach is the requirement of a normally distributed dataset. To overcome this issue of normal distribution, an alternative approach the ECOBWEB was proposed, that applies the feasibility assessment of the median variable rate (Gennari, Langley, & Fisher, 1989), (McKusick & Thompson, 1990), (Reich & Fenves, 1991).

A non-hierarchical unsupervised approach for mixed data was suggested. The benefit of this method was that it could handle data with missing values (Di Ciaccio, 2001). A method that applied the additive unsupervised method to mixed data was

suggested (Sowjanya & Shashi, 2011). Their approach involved a random cluster centre initialization. Then data were nominated to group basis of their similarity distance from each other within the group. To group categorical features the mode was used to compute similarity and the mean was used for the numeric features. It's unclear from the article on the choice of distance measure applied to form groups. However, it is not clear from the paper as to which distance measure was used to cluster the data points. The researchers proposed an affinity based propagation clustering algorithm (APC) that applied the concept of message transmission (Frey & Dueck, 2007). This method was extended by integrating the distance method proposed by Ahmad & Dey (2011) to the APC algorithm (Zhang & Gu, 2014). It was reported that they were able to attain good results using this approach. The researchers applied a divide and conquer approach to mixed data. The data was initially segregated into numeric and categorical features. Then they applied a graph based partitioning algorithm to group numeric features. The Squeezer algorithm proposed by He, Xu, & Deng (2002) was applied to group the categorical feature. Thereafter, the clustering results were collated and considered as categorical, to which the Squeezer algorithm was applied again to yield the final clusters. The loss of information during data conversion is not discussed in this paper (Elavarasi, Akilandeswari, & Sathiyabhama, 2011).

2.4.5. Key literature review observations

As outlined above, a majority of the clustering algorithms to process mixed data are partitional in nature, because these algorithms are:

- Easy to implement
- Linear with data objects
- Easy applicability to parallel architectures

Despite these advantages, determining a suitable similarity measure and rate objective to process mixed data is a challenge for partition-based clustering algorithms. The other clustering algorithms namely mode-based, neural network-based or even hierarchical unsupervised methods outshine other approaches, yet they deteriorate with non-determinate time or space convolution or invite data distribution supposition. This might impede their usability to real-world applications.

2.5. An overview of distance measures

The similarity of objects within a cluster play a fundamental role in the clustering process. A good cluster is determined by objects having maximum similarity between them. The measure of similarity in a cluster is evaluated in terms of distance between the objects. In conventional or *hard clustering*, an object will *absolutely* belong to a cluster or not. Often in practice, the distance between objects x_i and x_j in a cluster is denoted as; $d(x_i, x_j)$ where d is the distance between the objects. A valid distance measure must be symmetric i.e., $d(x_i, x_j) = d(x_j, x_i)$ and has a minimum value of zero in case of identical objects. Shorter the distance between any two objects; closer these objects are assumed in terms of similarity.

2.5.1. Distance measures for numerical data

Often, the numerical variable is discretized. For instance, the range of a numerical variable is split into a certain number of intervals. But such a discretisation raises two problems, namely, (a) determination of range intervals is imprecise and (b) the discretisation process can cause information loss (Hennig, Meila, Murtagh, & Rocci, 2015). This ambiguity coupled with clustering render it difficult to validate the choice of discretisation. This section has outlined distance measures for numerical data types as shown in Table 2.3.

Table 2.3: Distance measures for numerical data

No	Distance name	Functionality	Advantage	Disadvantage	Reference
1	Minkowski distance	Sensitive to outliers	Works well if the variables are isolated. Independent of the underlying data	Sensitive to outliers	(Irani, Pise, & Phatak, 2016)
2	Manhattan distance (is Minkowski distance of order 1)	Sensitive to outliers	Is a modified version of Minkowski Independent of the underlying data	Sensitive to outliers	(Irani et al., 2016)
3	Euclidean distance (is Minkowski distance of order 2)	Easy to understand and compute	Is a modified version of Minkowski Independent of the underlying data	Sensitive to outliers. Large scale value dominates others. Normalization is the solution.	(Irani et al., 2016)
4	Average distance	Sensitive to outliers	It is a modified version of Euclidean distance.	Sensitive to outliers	(Irani et al., 2016)
5	Pearson correlation coefficient	Can take a range of values from +1 to -1. Where 0 indicates no association, +1 & -1 indicates positive & negative associations	Measures the linear correlation between features.	Sensitive to outliers. Large scale value dominates others. Normalization is the solution.	(Irani et al., 2016)

2.5.2. Distance measures for categorical data

The mechanism involved in clustering categorical variables resembles to clustering numerical variables. However, the difference is in distance measurement for categorical variables where a matching based distance function is used. Some clustering algorithms implement a similarity matrix type data structure for measuring similarity between categorical variables. Some of the approaches followed for calculating distance between categorical variables are:

a. Conversion of categorical data variable into numeric and then applying numeric clustering algorithm like K-means.
b. Conversion of numeric data variable into categorical and then applying categorical clustering algorithm like K-modes.
c. Directly handling the mixed data.

The Table 2.4 presents the distance measures for categorical data.

Table 2.4: Distance measures for categorical data

No	Distance name	Functionality	Advantage	Disadvantage	Reference
1	Hamming distance	Distance between different categorical values is set at 1, while a distance of 0 is set for identical values.	Easy to understand	The main drawback is all feature values are considered as equal ignoring the statistical properties of the feature values.	(Norouzi, Fleet, & Salakhutdinov, 2012)
2	Pearson's chi-square statistic	Used to measure the separation amongst the categorical and numerical variables in a probability table.	Robust to data distribution. Permits evaluation of both dichotomous independent variables	It is a significance statistic measure. Difficult to interpret when there are large number of categories (20 or more)	(Sharpe, 2015)
3	Goodall's similarity	Uses probability as a normalization process to measure the likeness across objects	Ascribes a high similarity score if the feature values are less frequent than if the value is frequent	Computationally expensive process	(Boriah et al., 2008)
4	Anderberg similarity	Another probability-based measure. Assigns a weighting method to categorical features. The range is between [0,1]	High score (*near to 1*) is assigned to rare matches and lower similarity to rare mismatches	Computationally expensive process	(Boriah et al., 2008)

5	Lin similarity	An information theoretic measure.	The Lin measure ascribes increase weight to recurrent values, and decreased weight to sparsely occurring values.	Computationally expensive process	(Boriah et al., 2008)

Some other distance measures for categorical variable, in particular the nominal variable are of the following;

2.5.2.1. Simple Matching Coefficient (SMC)

The SMC also known as the overlap measure or the Rand similarity coefficient (Sulc, 2014). It is the simplest method to measure similarity. It should be noted that there is no difference between similarity measure and distance measure. When determining the likeness across variables m_c and m_d for the n object, it assigns a value 1 if the variables match, and a value 0 otherwise, of the *i-th* object as described by the formula

$$S_i = (m_{ci}, m_{di}) = \begin{cases} 1 & if\ m_{ci} = m_{di} \\ 0 & otherwise \end{cases}$$

The likeliness across two variables is measured as

$$(m_c, m_d) = \sum_{1}^{n} S_i = \frac{(m_{ci}, m_{di})}{n}$$

To create a proximity matrix, the dissimilarity between variables has to be computed. This leads to the overlap measure that is defined as $D(m_c, m_d) = 1 - S_i \sum_1^n S_i$ (Sulc, 2014). The overlap measure is essentially a similarity measure that determines whether two observations match or not. It should be noted that it does not consider frequency distribution of categories that could serve as an imperative factor in determining the

variable association. In literature, there exist other similarity measures that have tried to overcome this drawback.

2.5.2.2. Inverse Occurrence Frequency (IOF)

The IOF measure was originally developed for text mining adjusted for categorical variables (Sparck-Jones, 1972). This measure ascribes a higher similarity score to dissimilarity on less frequent values. For the i object, it is described as

$$S_i = (m_{ci}, m_{di}) = \begin{cases} 1 & if\ m_{ci} = m_{di} \\ \dfrac{1}{1 + lnf(m_{ci}).lnf(m_{di})} & otherwise \end{cases}$$

where $f(m_{ci})$ expresses the frequency of the category m_{ci} of the $i - th$ object. In stark contrast to IOF is the Occurrence Frequency (OF) measure that assigns lower similarity to mismatches on less frequent values and is given as

$$S_i(m_{ci}, m_{di}) = \begin{cases} 1 & if\ m_{ci} = m_{di} \\ \dfrac{1}{1 + ln\, f(m_{ci}).lnf(m_{di})} & otherwise \end{cases}$$

2.5.2.3. Lin measure

This measure was introduced by Lin (1998). It represents the information theoretic similarity based on relative frequencies. It assigns higher similarity to most frequent matching categories and lower similarity to the least frequent mismatching categories. In equation form, it's expressed as

$$S_i(m_{ci}, m_{di}) = \begin{cases} 2\ ln\, p(m_{ci}) & if\ m_{ci} = m_{di} \\ 2\ ln\left(p(m_{ci}) + p(m_{di})\right) & otherwise \end{cases}$$

Where $p(m_{ci})$ expresses a relative frequency of the category m_{ci} of the $i - th$ object.

2.5.3. Distance measures for mixed data

Rather than recoding categorical or numerical variables, an alternative approach is to define a dissimilarity measure for each type of variable which are then combined

together. In 1971, Gower proposed a mathematical formulae of distance measurements for mixed data (Gower, 1971). This finding is important to mention as it has established a foundation for measuring nominal feature. The Gower algorithm determines similarity between features by applying the Manhattan distance for numerical features. According to this algorithm, assume a data matrix $A = \{a_{xy}\}$ where $x = 1,2, \dots, n$ (the number of features is denoted by$n$) and $y = \{1,2, \dots, f\}$ ($f$is the number of features). Then, the dissimilarity between the objects $a_x = [a_{x1}, a_{x2}, \dots, a_{xn}]$ and $a_y = [a_{y1}, a_{y2}, \dots, a_{yn}]$ is expressed by the formula $d_H(a_x, a_y) = \sum_{y=1}^{f} d_{xyf}$ where d_{xyf} is a similarity measure between $x - th$ and $y - th$ objects by the $f - th$ variable. This formula will only work for datasets with complete entries. Besides, this formula considers a nominal feature to have only two categories, i.e. if given two nominal features match, the digit 0 is then assigned, and when the categories do not match, the digit 1 is assigned. The numeric features are range normalised. The ordinal features are rank-ordered and subtracted by 1 and finally range-normalised like the numeric features.

The Gower's similarity coefficient formula is given in equation (1),

$$S_{ij} = \frac{\sum_{k=1}^{n} s_{ijk}\delta_{ijk}w_k}{\sum_{k=1}^{n} \delta_{ijk}w_k} \quad (1)$$

Where the number of features is denoted by n, s_{ijk} is the likeness across i and $j$measured on the k^{th} feature δ_{ijk}equates to null if value of the k^{th} feature is missing for either of the two objects $i$and$j$, and is 1 if it's available for both objects, and w_k is the feature weights. This is a simplistic approach. The equation (1) was modified and is discussed in sub-section 4.5.3

In 1997, the K-prototypes method for mixed data clustering was suggested (Huang, 1997). It included three phases; the *elementary group selection*, *cluster appropriation,* and finally the *re-allocation*. In the initial step, a randomised selection of *n* data points as cluster centres were made. Next, squared Euclidean distance metric was applied to compute similarity between features of numerical types. Thereafter, distance measurement for categorical features is based on their mode.

And finally, in the third step which is the reallocation phase, the inceptive group centroids obtained in step 1 & 2 are recalculated until a local optimum is reached. The algorithm computing cost is $O((t+1)kn)$ and $n$is the count of data values, $k$initial count of groups, $t$is the iteration sequence of the reallocation process. Huang's algorithm has a major shortcoming in the selection of features. Notably, for numerical values the distance measured is squared Euclidean distance, which is susceptible to high values, whereas for categorical features the distance measured is frequency. Only high frequency valued categorical features are considered by this algorithm. The categorical features with a lower frequency are discarded that directly leads to information loss. The proposed approach for mixed data analysis is presented in sub-section 4.3. Since Gower proposed this algorithm, there have been several improvements to this approach. In 1999, Podani extended Gower's general coefficient similarity work for ordinal features. Podani argued that in the Gower's method for ordinal feature treatment, there was a loss of information in data conversion (Podani, 1999). To overcome this problem, Podani suggested to initially rank order all ordinal features. The features with similar rank are close to each other and do not influence the results. Then count the number of steps between similar rank features and other features. In essence, Podani's approach is similar to nearest neighbour classification approach in a "partial [rank] order". In the year 2006, a group of researchers proposed a weight-based approach to remedy the problem associated with Gower's approach. Their idea was by assigning weights will prevent feature dominance. To measure feature similarity for numeric features, Pearson's correlation coefficient was used and for categorical features, product moment correlation was applied. The categorical features were binary encoded. However, it is not clear the data type of categorical feature, if it was nominal or ordinal. Moreover, the authors applied principal component analysis to reduce data dimensionality for obtaining significant features (Chae et al., 2006). Continuing further, in particular this thesis discusses a recent paper that presented three modifications for treating nominal data. In the first modification, the

authors introduce *"variable entropy"*, in which the concept of weights is used. A higher weight is given to a nominal feature that has a higher variability. The authors assert that such variables are rare as compared to nominal features with lower variability. Although, logically this assertion is incorrect because nominal features with higher variability are actually not *rare* but rather prominent and thus assigned higher weights (Šulc, Matejka, & Procházka, 2016). The inference is coherent with the findings on assignment of weights to nominal features (Gower, 1971). In the second modification, the authors applied the *"Inverse Occurrence Frequency"* concept and assigned higher weights to infrequent mismatches between the nominal features. In the third modification, the authors assigned higher weights to mismatches between nominal features having a smaller number of categories (Šulc et al., 2016). This approach was also comparable to the method recommended by (Lin, 1998). This assigned weight takes a value between 0 and 1. It is important to note, that like Gower, the authors proposed modifications can only work for two categorical levels. Also, the authors had not discussed the numeric feature treatment in their proposed modifications unlike the Gower's method where numeric feature treatment was elucidated. Furthermore, the authors tested their proposed modifications for hierarchical clustering method, namely the two-step cluster analysis and the latent class analysis and Rand Index, an external cluster validation metric was used. A comprehensive review on distance methods for mixed data clustering is presented (Velden, D'Enza, & Markos, 2018). The Table 2.5 details popular distance measures for mixed data.

Table 2.5: Distance measures for mixed data

No	Distance name	Functionality	Advantage	Disadvantage	Reference
1	Gower distance	The features are divided into two subsets; categorical and numerical. It uses a range-normalized Manhattan distance for numerical data. The categorical data are initially transformed into m-two factored variables and the Dice coefficient is applied.	Intuitive to understand and easy to compute	Categorical data conversion incurs information loss. Sensitive to outliers present in numerical data.	(Gower, 1971)

2.6. An overview of unsupervised feature selection approaches

In literature there exist both supervised and unsupervised feature selection approaches. There are three approaches to perform FS for unlabelled data, namely, the filter, wrapper and hybrid approach (Solorio-Fernández, Martínez-Trinidad, & Carrasco-Ochoa, 2017).

Some well-known feature selection approaches in literature are the Principal Component Analysis (PCA), Correspondence Analysis (CA) and Multiple Factor Analysis (MFA) (Abdi & Williams, 2010). Mathematically, PCA depends upon the Eigen-decomposition of positive semi-definite matrices and the singular value decomposition (SVD) of rectangular matrices. Researchers have suggested that PCA is a feature extraction algorithm and not feature selection because it transforms the original feature set into a subset of interrelated transformed features, which are difficult to emulate (Abdi & Williams, 2010). Moreover, PCA only works for numerical data. The algorithms CA and MFA are generalizations of PCA where CA can handle categorical data and MFA can handle mixed data. And since PCA and its variants transform the original feature set,

therefore, in this thesis they will not be considered for algorithm performance comparison, which is detailed in chapter 5.

This thesis is focussed on unsupervised filter-based feature selection; therefore, a brief overview of the existing unsupervised filter-based feature selection methods only is discussed in this section. Filter methods essentially, use the intrinsic properties of the data to select features. In Table 2.6, the existing filter based unsupervised FS methods is given. The researchers Dash, Liu & Yao (1997) suggested a univariate unsupervised filter based feature selection method that applied "entropy of similarities". This entropy was computed by the total entropy of a similarity matrix W, where the data points of W were the similarity pair of data points in the dataset. The similarity in W was calculated as follows: if all the data points in the dataset were numerical, the similarity between them was computed by applying the Euclidean distance exponential function; conversely, if all the data points in the dataset were categorical in nature, then the similarity between the data points was computed by the Hamming distance. In contrast, if the dataset consisted of mixed data, the researchers suggested to discretize the numerical data points into categorical before applying the Hamming distance. The feature relevancy was determined by computing the "leave-one-out" sequential backward process, which was then combined with the entropy measure outlined earlier. The final result was a ranked feature matrix consisting of the most relevant features first followed by the lesser relevant data points. It is worthy to mention the research works related to unsupervised filter-based feature selection methods for the numerical data. According to the researchers Niijima & Okuno (2008), off all the univariate filter methods for unsupervised feature selection, the two most notable ones are the Singular Value Decomposition (SVD) Entropy method (Varshavsky *et al.,* 2006) and the Laplacian Score method (He, Cai & Niyogi, 2006). The SVD Entropy method computes the relevance of data points in a dataset by calculating their dissipated entropy which is based on the SVD matrix (Alter, Brown & Botstein,

2000). On the contrary, for the Laplacian Score method, the fundamental idea is to determine the feature ranking which is strongly dependent on their neighbourhood power computed by the Laplacian graph (Belkin & Niyogi, 2002). Yet another pertinent univariate filter-based feature selection approach is SPEC (Zhao & Liu, 2007). It computes feature ranking by applying the Eigen-system present in the Laplacian matrix which contains the computed similarity distance between the data points. Mitra et al. (2002), suggested a multivariate filter approach called the Feature Selection using Feature Similarity (FSFS). This approach functions by taking into account the dependency or the similarity distance between the data points in a dataset. The similarity is computed basis of the accountable variance-covariance among the data points. Their proposed approach initiates by dividing the initial dataset into groups such that data points within the groups are similar to each other and dissimilar to other data points in other groups. It then selects one representative data point from each group, which is then included into the final data subset.

Table 2.6: Filter based unsupervised feature selection methods for mixed data clustering

S. No.	Feature Selection (FS) approach	Reference
1	Spectral clustering-based FS approach is used. Spectral clustering is based on graph theory. The similarity is computed by analysing the spectral gap of the normalized Laplacian matrix.	(Solorio-Fernández, Martínez-Trinidad, & Carrasco-Ochoa, 2017)
2	A univariate filter approach called "Sequential Backward Selection for Unsupervised Data" is used.	(Dash, M., Liu, H., & Yao, J., 1997)

The unsupervised filter-based feature selection methods may further be classified into two categories defined for univariate and multivariate data. The univariate methods apply feature relevance measurement using an external criterion. There are numerous well-known unsupervised univariate filter-based feature selection approaches, like, “information gain”, “gain ratio”, “symmetrical uncertainty”, “Gini index” and “Fischer score”. Furthermore, the difference between univariate and multivariate unsupervised filter-based feature selection methods is, in the former dependency between the features are ignored, whereas in the latter, the dependencies between the features are considered to evaluate the relevant features. Thus, the multivariate methods are computationally more expensive as compared to the univariate methods, although their performance is better than the univariate methods. Some examples of multivariate feature selection methods are, “minimal-redundancy-maximal-relevance (mRMR)”, “mutual correlation”, “random subspace method”. The PCA is a multivariate unsupervised filter based feature selection method. Similarly, “Bhattacharya distance”, “Wilcoxon Paired Test”, “ROC-based test”, “Entropy-based test” and “Laplacian score”, are unsupervised filter based feature selection methods for the univariate data. The Laplacian score determine the feature importance by calculating the feature’s ability to preserve the local distance. So while, the unsupervised feature selection is more challenging because of the absence of labelled data, it has still several advantages associated to it. For instance, it’s unbiased as it does not require experts or data analysts to classify the samples. And it can still work well when no prior information is available. Besides this, it’s also helpful in exploratory data analysis as it provides an effective way to determine the unknown pertinent insights from the dataset. The main drawback with unsupervised filter-based feature selection methods are that it ignores the interaction between features or better known as correlation in supervised parlance. And this, the possible interaction among features (including the combined feature set may portray an effect which necessarily is not reflected by

individual features in a group). This may yield varied cluster purity results when the same feature set is applied to differing clustering algorithms.

2.7. An overview of data transformation approaches

In this section, a discussion on existing approaches on mixed data transformation for clustering is presented. First, it will discuss the discretisation from numeric feature to categorical and the use of appropriate categorical clustering method. Then, a discussion on numerical coding of categorical features for clustering is presented.

2.7.1. Discretisation

Discretisation of a numerical feature is a widely used method in statistics and machine learning. In this approach, all numeric features are discretised and an applicable clustering method for categorical data is used (e.g. the K-modes algorithm (Huang, 1998)). A possibility of data loss is imminent in the discretisation process if inappropriate cut-off points are used (Foss et al., 2018). Another reason attributed to data discretisation is to transform the numeric data into categorical data, often in the form of discrete or nominal variables with a finite interval set. In practice, data discretisation can be perceived as a data reduction method Garcia et al (2012). A well-designed survey on data discretisation methods in supervised learning is presented by Garcia et al (2012).

2.7.2. Numerical coding

In this approach, the categorical data is transformed to numeric and an appropriate clustering method is applied like the K-means algorithm. Often direct replacement is not possible so other methods like dummy coding and simplex coding are used (Foss et al., 2018). In practice, clustering with numeric coding always involves applying a 0-1 dummy coding with standardized numeric features. Researchers have shown that this strategy is not conducive for an equitable balancing of numeric and categorical features for

clustering process. Yet another approach is to assign suitable weights to categorical features and then perform clustering. This approach may work for certain environmental settings however, is not applicable in a general sense (Foss et al., 2018).

2.8. Cluster Validation Metric (CVM)

While the cluster development is an important process, it's equally important to test the accuracy and validity of the cluster. The clusters obtained through a certain method must be evaluated, on parameters such as maximum similarity between the objects within the cluster, and minimum similarity with objects from other clusters recently, many evaluation criterions have been developed. Verifying the validity of a clustering process is an arduous task and there is a paucity in literature as enjoyed by the classifier algorithms. Although previous studies have shown that there is no single CVM that outshines the rest (Zhao, Liang, & Dang, 2017). Nevertheless, it is important to outline CVM methods. There are three types of CVM, internal, external and relative validation but the classification criteria is not always clear (Halkidi, Batistakis, & Vazirgiannis, 2001), (Jain & Dubes, 1988), (Brun et al., 2007) and (Pfitzner, Leibbrandt, & Powers, 2009). Nevertheless, there is a clear distinction between the CVM if the focus is on the information present in the validation process. In literature, there have been extensive surveys that have detailed the various types of CVM. For instance, the paper published in 1985 is still the work of reference on internal cluster validation (Milligan & Cooper, 1985). That work compared 30 cluster validation indices. The paper by Milligan & Cooper was further refined (Arbelaitz, Gurrutxaga, Muguerza, PéRez, & Perona, 2013) who compared 30 CVM in 720 synthetic and real-world datasets. They have categorized Silhouette Coefficient (SC) as an internal CVM and Accuracy as a classification validation metric Arbelaitz et al (2013). This thesis has used the Arbelaitz et al (2013) paper as the benchmark to justify the choice of the internal CVM, the Silhouette Coefficient (SC). The internal CVM like SC, Dunn Index (DI), and Davies-Bouldin Index (DBI) rely upon the internal clustering information of the process without referencing any external information. It should be noted that these internal CVM are applied when the ground truth

labels (or in a classification parlance, the class labels) are unknown Arbelaitz et al (2013). Other methods related to external validation like Accuracy, Rand Index (RI), Adjusted Rand Index (ARI), Jaccard, Fowlkes-Mallows and Callinski-Harabaz Index (CHI) also known as variation of information criterion-evaluate a cluster division by a comparison with an already known correct partition, or in other words, when the ground truth labels are known then these external CVM are applied Arbelaitz et al (2013). The relative CVM evaluate the cluster by exercising varied parameter values in an algorithm (e.g., several reiterations of the cluster numbers) Arbelaitz et al (2013). The focus of this thesis is on unsupervised approach, therefore, a discussion to internal CVM is apt and is shown in Table 2.7.

Table 2.7: Internal Cluster Validation Metrics

No	Evaluation metric	Functionality	Advantage	Disadvantage	Reference
1	Silhouette Coefficient (SC)	Measures the degree of confidence in clustering assignments. Well clustered objects are given a value near 1 and poorly clustered objects are given a value near -1. A score around 0 indicates overlapping clusters.	The score is higher when clusters are dense and well separated.	The silhouette score is generally higher for convex clusters like the one's obtained through density-based clustering.	(Rousseeuw, 1987)
2	Calinski-Harabaz Index (CHI)	For *k* clusters, it gives a ratio between cluster dispersion mean and the within cluster dispersion	Is fast to compute. The score is higher when clusters are dense and well separated.	The CHI score is generally higher for convex clusters like the one's obtained through density-based clustering.	(Caliński & Harabasz, 1974)

3	Davies-Bouldin Index (DBI)	It measures the average similarity between each cluster C_i and the most similar cluster to it C_j Values close to zero indicate a good partition.	The computation is simpler than SC	The DBI score is generally higher for convex clusters like the one's obtained through density-based clustering. A good value reported by this metric does not refer to an optimal information retrieval.	(Halkidi, Batistakis, & Vazirgiannis, 2001)

2.9. Educational Data Mining (EDM)

The EDM approach involves the translation of raw educational data points generated from educational systems into useful information that can help its handlers to serve their customers well (Romero & Ventura, 2010). Using classical data mining approaches like classification, clustering have given researchers good results in the past (Mohsin, Norwawi, Hibadullah, & Wahab, 2010).

2.9.1. Clustering algorithms applied in EDM

A homogenous dataset consists of univariate data type, whereas a heterogeneous data consists of multivariate or both numerical and categorical data types. There are innumerable studies in the literature that have addressed numeric datasets inherent in educational contexts. In Table 2.8, has described these studies succinctly. This thesis has found scarce research on this subject.

Table 2.8: Clustering methods adapted in EDM

No	Reference	Problem/ Objective	Algorithm	Dataset/Data source	Group	Datatype
1	(Chen & Cooper, 2001)	To detect & mine student web usage pattern	FASTCLUS, Ward's	Transaction log data of an LMS	learning style	Numeric
2	(Rashid et al., 2011)	To classify student learning style	Two-step cluster analysis	Dataset given by Sultan Idris Education University, Malaysia	learning style	Numeric
3	(Zheng, Du, & Tian, 2007)	To determine student groups	K-means, Farthest First & Expectation-Maximisation (EM)	LMS data from Xi'an Jiaotong University	e-learning	Numeric
4	(Dradilova, Martinovic, Slaninova, & Snasel, 2008)	To visualise student learning patterns	Hierarchical clustering	LMS data from Silesian University	e-learning	Categoric
5	(Feng, Shibin, Cheng, & Qinghua, 2008)	To analyse student learning in e-learning environment	Fuzzy C means, K-means	Dataset from Xi'an Jiaotong University"	e-learning	Numeric

6	(Jili, Kebin, Feng, & Huixia, 2009)	To group the learner behavioural patterns in a e-learning environment	Fuzzy clustering	A qualitative method is proposed	e-learning	Categoric
7	(Aher & Lobo, 2012)	To suggest an optimal course to learner	K-means, Apriori association rule	LMS data	e-learning	Numeric
8	(Antonenko, Toy, & Niederhauser, 2012)	To group learner behavioural patterns in a e-learning environment	Ward's and K-means	Dataset of 59 students from a Mid-Western University	e-learning	Numeric
9	(Cobo et al., 2012)	To group learner behavioural patterns in a e-learning environment	Hierarchical clustering	LMS data	e-learning	Not given
10	(Eranki & Moudgalya, 2012)	To group learner behavioural patterns in a e-	K-means	LMS data	e-learning	Numeric

		learning environment				
11	(Ghorbani & Montazer, 2012)	To group learners on their emotional learning style	K-means, C-means, evolutionary fuzzy	LMS data	e-learning	Numeric
12	(Valsamidis, Kontogiannis, Kazanidis, Theodosiou, & Karakos, 2012)	To analyse LMS data	Markov Clustering, Simple K-means	LMS data from Technology Education Institute, Kevala	e-learning	Numeric
13	(Romero, López, & Ventura, 2013)	To determine relevant features for predictive modelling	EM, Hierarchical Cluster, SIB, K-means	Dataset of semaphore year in information technology	e-learning	Numeric
14	(Chen, Chen, & Liu, 2007)	To identify learning performance assessment rules	Gray correlated theory, K-means, fuzzy conjecture	Student assessment dataset	e-learning	Numeric

15	(Manikandan, Meenakshi Sundaram, & Mahesh Babu, 2006)	To group students with similar learning preferences	K-means	Student learning dataset	collaborative learning	Numeric
16	(Anaya & Boticario, 2009)	To group learner behavioural patterns in a e-learning environment	EM	Student learning dataset from UNED European universities	collaborative learning	Numeric
17	(Huang, Lin, Wang, & Wang, 2009)	To determine relevant features for predictive modelling	Cluster Analysis, Linkage Method	Student learning dataset of China Motor Corporation.	collaborative learning	Numeric
18	(Chang, Wang, & Li, 2010)	To group learner behavioural patterns in a e-learning environment	Item-Response theory & K-means	Student learning dataset	collaborative learning	Numeric

19	(Wook et al., 2009)	To build predictive models from historical educational data	ANN, Farthest First, Decision Tree	Student learning dataset	EDM	Categoric
20	(Salazar, Gosalbez, Bosch, Miralles, & Vergara, 2004)	To determine relevant features for predictive modelling	C-means	Student learning dataset from Industrial University of Santander	EDM	Numeric
21	(Dharmarajan & Velmurugan, 2013)	To build predictive models from historical educational data	CHAID classifier	Student assessment dataset	exam failure	Numeric
22	(Almeda, Scupelli, Baker, Weber, & Fisher, 2014)	To determine the classroom wall decoration style by teachers	K-means	Student learning dataset	classroom decoration	Numeric

23	(Ivancevic, Celikovic, & Lukovic, 2012)	To determine reasons for student seating choice & its impact on assessments	K-means	Student learning dataset	learner seating order	Numeric
24	(Chen, Chan & Lin, 2007)	To group learner behavioural patterns from e-learning environment	EM, K-means	Transaction log data of an LMS	learning portfolio	Numeric
25	(Li & Yoo, 2006)	To group learner behavioural patterns in a e-learning environment	C4.5, Bayesian Markov Chain	Transaction log data of an LMS	Student modelling	Categoric
26	(Baker & Gowda, 2012)	To determine factors responsible for shallow learning	Statistical measures	Transaction log data of an LMS	Student modelling	Not given

27	(Chi, Kuo, Lu, & Tsao, 2008)	To build predictive models from historical educational data	Hierarchical K-means	Transaction log data of an LMS	profiling clustering	Numeric
28	(Trandafili, Kajo & Xhuvani, 2012)	To build predictive models from historical educational data	EM, association-rule and decision tree	Student learning dataset	profiling clustering	Categoric
29	(Tair & El-Halees, 2012)	To build predictive models from historical educational data	Lift-metric, Rule-based, Naïve Bayesian, K-means	Student learning dataset	student performance	Numeric
30	(Bharti, Shukla, & Jain, 2010)	To determine factors responsible to curb class dominance	K-means	Student learning dataset	intrusion detection	Numeric

31	(Cobo et al., 2011)	To group learner behavioural patterns in a e-learning environment	Agglomerative Hierarchical clustering.	Student learning dataset	learner behaviour	Categoric
32	(Perera, Koprinska, Yacef, & Zaiane, 2009)	To group learner behavioural patterns in a e-learning environment	K-means & EM	Transaction log data of an LMS	Computer supported collaborative learning	Numeric
33	(Amershi & Conati, 2009)	To group learner behavioural patterns in a e-learning environment	K-means	Transaction log data of an LMS	Student modelling	Numeric
34	(Sardareh, Aghabozorgi, & Dutt, 2014)	To aid the importance of reflective dialogues in student learning	Hierarchical clustering and K-means	Student learning dataset	classroom learning	Numeric

The undergraduate students' academic performance was evaluated in a study that suggested to integrate several disparate data drilling approaches such as Farthest-First method dependent on *K*-means clustering, Artificial Neural Network (ANN), and Decision Tree into a unified method. This approach was then applied to an educational dataset that was sourced from an educational institution in Malaysia (Wook et al., 2009). The researchers claim to have drastically augmented the existing K-means clustering approach which suffered from several impediments. In their study, the researchers further wrote the existing k-means approach was susceptible to the cluster centre initialization issue, it will fail to converge to a local optimal value. Besides these issues, the other problem was the high time complexity when dealing with high dimensional datasets (Zheng & Jia, 2011).. So to curtail these drawbacks, the researchers proposed an improved formulation of the K-means clustering approach (Zheng & Jia, 2011).

In a research study the applications of various DM approaches were applied educational datasets. The Apriori algorithm was subjected to student dataset such as to derive the most appropriate and meaningful data association rules that further helped in augmented profiling of student activities. Incidentally, the K-means approach was utilized to group data. Although the research stated to have obtained the dataset from an educational institution but it did not state anything about the database holding the dataset (Parack, Zahid, & Merchant, 2012b). The research conducted by (Zhiming & Xiaoli, 2008) on undergraduate students of a university found the occurrence of significant data points that can be leverage to augment student performance. The researchers applied the C-means grouping approach but it did not state much about the data points and its properties. In another research study (Zheng et al., 2007), the K-means approach was applied to group high dimensional educational dataset to curtail the high computation complexity. The researchers suggested a new approach that applied the Co-operative Particle Swarm Optimizer (PSO). In order to study the student profile formed when

students interact with an e-learning system a study was conducted (Chi, Kuo, Lu, & Tsao, 2008). The underlying idea was two-fold: primarily, the suggested approach filtered the learning material basis of its contents to extract relevant keywords, quite similar to text mining. The second step, involved the application of the hierarchical k-means to this bag of words obtained from the first step. This way the webpages browsed by the student activities were filtered and the students were then recommended appropriate web pages basis of their browsing patterns.

When students learn in a collaborative environment using an e-learning platform then their online traces like browsing patterns can help in building their portfolio's. In an e-learning parlance, such portfolios are named as "learning portfolio". And such portfolios can also be created in an offline learning environment by utilising the student collaborative activities like group discussion or group project working. In a very specific study related to this topic, the researchers Chen et al. (2007) integrated the K-means, EM and Farthest First clustering approaches with *t*-test to the student portfolios present in an distance learning based information system. Using the unsupervised approach as the central idea of their research study, they uncovered several interesting facts from the student leaning patterns. Specifically, the student t-test method was applied to compute the mid-term student performance with their final-term examination performance, and also evaluated the groups comprising high and low learning performances in exams. The dataset was sourced from an educational institution. The researcher's experimental study found the presence of a positive correlation between students who attended all classes and had also obtained significantly better scores. The researchers also found that there existed no correlation between the student online browsing pattern and the duration of time spent in completing an e-learning assignment. However, the study reported that student performed significantly better in exams if they had discussed the course material using the e-learning environment.

Another similar study related to studying the effects of e-learning environment called the TRAC, amongst the students was conducted (Perera et al., 2009). The K-means as the unsupervised approach & EM algorithm that are integrated in the WEKA software were used to determine groups. Furthermore, the researchers applied the hierarchical agglomerative grouping and chose the Euclidean distance for similarity measure. The students were required to use this system for any sort of online learning including collaborative learning. Using this e-learning information system, the data for three semesters was collected and used for analysis. The researcher's unravelled important findings especially the ones related to student collaboratively learning in an e-learning environment. Such instruments of learning helped to foster team building opportunities among the students. From the educator's perspective, this environment helped them better understand the complex tools that were used for improving the learning and teaching efficiency of both the educator and the learner in a simple manner.

2.9.2. Mixed data clustering approach in EDM

There is scarce research which studied the mixed data clustering in EDM. This work discovered one particular study which had addressed this issue. The researchers (Shuangyan Liu & d'Aquin, 2017) applied the K-prototypes algorithm to a mixed data consisting of student demographics (categorical) and achievements (numerical) in a distance learning program. However, that study failed to discuss its feature selection. In this paper, the authors filled the missing values with zeroes. The Elbow method (Kaufman & Rousseeuw, 2009) was used to determine the number of groups that were then passed into the K-prototypes algorithm. They wrote a custom program in Octave programming language to emulate the K-prototypes algorithm, claiming that they were unable to find the K-prototypes implementation in programming languages such as R or Matlab. There exist implementations of K-prototypes in programming languages such as R and Python (Szepannek, 2018). The authors neglected to report the validation of their results either.

Two experiments were designed to emulate the approach (Liu & d'Aquin, 2017). In the first experiment, the K-prototypes algorithm was applied directly to their dataset. In the second experiment, using the proposed feature selection approach was applied to determine relevant features. Thereafter the K-prototypes algorithm was applied. A comparative result of both experiments confirmed the necessity of feature selection approach. More research is required in this area. This further raises an important question, "*Is applying a clustering method a feasible approach, when previous studies have used simple inferential statistical methods?"* This thesis argues that clustering is a pre-processing method. It helps in detecting groups by studying the object properties which can further be evaluated. The Table 2.9 shows existing work done for mixed data clustering in EDM.

Table 2.9: Mixed data clustering in EDM

No	Reference	Problem/ Objective	Algorithm	Dataset/Data source	Data type
1.	(Liu & d'Aquin, 2017)	To determine student achievements	K-prototypes	Mixed dataset consisting of student demographic details and examination records	Mixed data

2.10. Chapter summary

In this chapter, a comprehensive discussion of previous literature in mixed data clustering is given. This subject was approached by providing an overview of distance measures for numerical and categorical data. Then in section 2.6, a detailed discussion of research studies on clustering algorithms applied in EDM is given. It also presented the disparity in research in sub-section 2.6.2. Continuing further, the Table 2.5 shows the distance measures for mixed data. In Table 2.6, the filter based unsupervised feature selection methods for mixed data clustering and in Table 2.8 the mixed data clustering in EDM are shown. These tables reveal a gap in literature with very few researches on mixed

data clustering in the EDM domain. The following chapter will elucidate the research methodology.

CHAPTER 3: RESEARCH METHODOLOGY

3.1. Introduction

This chapter explains the research methodology used in this work. The sub-topics in this chapter include the motivation for designing a mixed data clustering approach for numerical and nominal data. In addition, the approach used for evaluation of the model and methods are presented. The last section concludes the chapter summary.

This work uses a school panel dataset of all primary schools in state of Delhi for academic session 2012-2013. An educational dataset consists of mixed data types. The proposed approach will accept an educational dataset and split it into categorical and numerical data frames. These data frames will then be pre-processed to eliminate issues such as collinearity, multicollinearity, outliers, missing data, skewness, and kurtosis to yield statistically significant variables. Next, these variables will be checked for clustering tendency using Hopkins statistic, whereby variables exhibiting the tendency will be retained. Finally, the variables will be grouped together and passed into a partitional clustering algorithm to yield pure clusters.

3.2. Research approach

Unlike the existing literature reviewed in the previous chapter; this research proposes an alternative approach which utilises statistical pre-processing techniques as a preliminary step for a partitional clustering algorithm. Besides considering the numerical variables inherent in educational datasets like attendance or examination records which have been applied by most of the research work as discussed in chapter 2, this research also considered categorical variables like mid-day meals and school locations. The method used for the proposed approach include steps that are depicted in Figure 3.1. The following sub-sections present the detail of each stage.

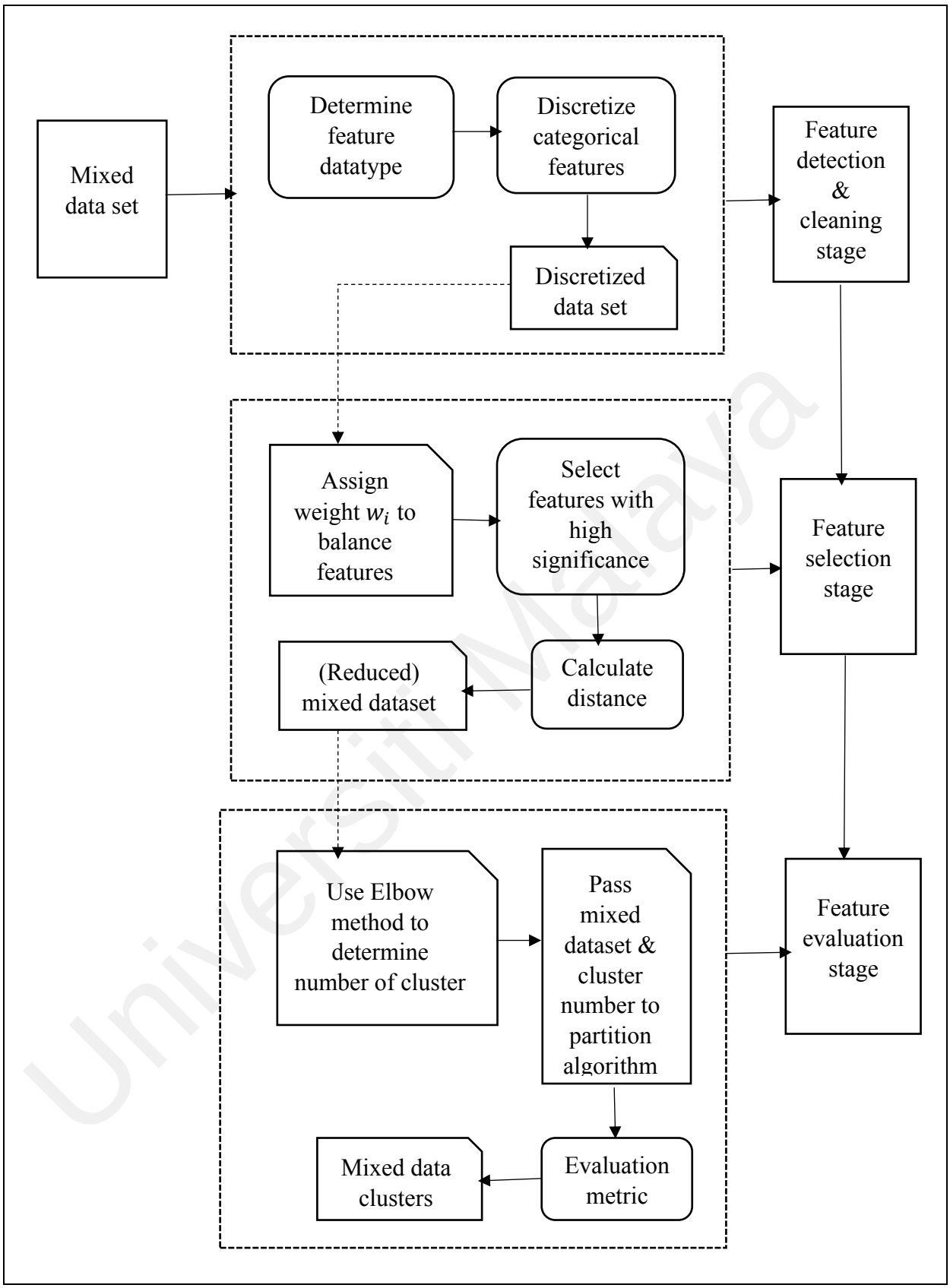

Figure 3.1: Proposed feature clustering approach

3.3. Research methodology

A typical educational dataset consists of categorical and numerical features (*a mixed dataset*). A numerical feature like student exam score or lecturer class hours will contain a numeric value. Similarly, a categorical feature like school location, will contain a text value. The proposed approach will accept educational data as an input and will yield cluster of similar features. In the first stage, the proposed approach will determine the nature of input data. It will then be segregated it into categorical and numerical features. Both type of features will be treated for issues such as missing data treatment, collinearity and multicollinearity, correlation, skewness, near zero variance as well as outliers. The idea is to conduct rigorous data pre-processing such that only the statistically relevant features remain in the data. It will also help in data dimensionality reduction. This subset will then be checked for clustering tendency. If the subset is group-able, the algorithm will show the possible number of groups. If the subset is not group-able the algorithm will stop. By the end of first stage, a subset of the original features is obtained. This subset will contain only the important features and the subsequent clustering will become more efficient.

In the second stage, the number of possible groups from the first stage will then be passed into the proposed partition-based feature selection approach to obtain clusters. An appropriate distance metric will be chosen to evaluate the subset containing mixed data types.

In the third and final stage, the obtained clusters will be checked for cluster purity in terms of accuracy and the result will be evaluated with the baseline methods.

3.3.1. Feature detection stage

Once an educational dataset is input from the user, the proposed approach identifies the feature type to form a feature matrix. Rows of the matrix represent a feature, and columns denote the feature. The *detector* identifies the data value contained in the feature and basis of the data type it then splits the feature into the appropriate type like categorical or numerical data type.

3.3.2. Feature cleaning stage

By the end of stage 1, the original feature set is divided into categorical and/or numerical data frames. In this stage, a comprehensive pre-processing is conducted. For instance, the missing values in the categorical data frame are replaced with mode and median in the numerical data frame. The categorical data are treated for issues like correlation significance and effect size. The presence of highly correlated variables does not improve cluster purity because of redundancy. To eliminate this, it is to ensure if variables are unrelated or are non-redundant. Features that are highly correlated to each other are non-contributors to a clustering model. Because such features act as noise and pollute the cluster. Therefore, it's important to obtain features that are uncorrelated to each other. The feature can either be categorical or numerical in nature. To determine feature independence between categorical variable, this thesis applied the Spearman's correlation method. And to determine feature independence between numerical variables, this thesis applied the Pearson correlation method.

3.3.2.1. Results Normalisation

The representative features are then normalised to yield a final result. Since, each of the feature type (categorical or numerical) is in its normalized form, the approach uses a weighted method to aggregate the results given as,

$$Result_{normalised} = Categorical_{result} + Continuos_{result}/2$$

3.3.3. Feature selection stage

Feature selection (FS) is a method to determine features that make maximum contribution to a model. The focus of FS method is to determine high variance objects from the original set, dependent on several feature maximization criteria. Often FS is referred to as dimensionality reduction. The difference between FS and dimensionality reduction is, the former must be a subset of the original features while the latter reduces dimensionality by creating new synthetic features from the linear combination of the original feature set. For example, Principal Component Analysis (PCA) is an unsupervised dimensionality reduction method (Li et al., 2018). Often in literature, PCA is referred to as a feature extraction method because it creates new synthetic features from the existing feature set. However, interpretability of such extracted features is difficult. This thesis is particularly focussed on the unsupervised feature selection. The FS methods are categorized into various type;

3.3.3.1. Supervised Feature Selection (SFS) methods

SFS is specified for classification type problems. It works by detecting feature correlation with the class label. An SFS method when applied to a dataset, works as $D = (X, C)$, consisting of features $X = \{x_1, x_2, \dots, x_n\}$ and class label C. The model objective is to determine an optimum feature subset $|S^{\wedge}(|S^{\wedge}|k^{\wedge})|$ that yields maximum model accuracy (Cai, Luo, Wang, & Yang, 2018), (Li et al., 2018).

3.3.3.2. Unsupervised Feature Selection (UFS) methods

In cluster analysis, features are regarded as similar if they contain the similar structural information. In conducting feature selection, the objective is to determine a minimum number of features that exhibit maximum structural information. To this end, the selected features must be as dissimilar as possible. The algorithm proposed here is similar to hierarchical feature clustering except that it forces every cluster to contain similar features. The likeness across two features is

determined using the absolute value of correlation. The likeness across the two clusters is then defined as the minimum similarity between each element of one cluster and the elements of the other. Once the clusters are formed, the features that do not comply with any cluster are eliminated.

3.3.4. Feature evaluation stage or Cluster validation metric

In a clustering process, there are no predefined classes and no examples that would demonstrate relations among data that is why it is perceived as an unsupervised process. On the other hand, classification is a process of assigning a data item to a predefined set of categories. Clustering produces initial categories in which values of a dataset are classified during the classification process (Halkidi et al., 2001)

To determine the validity of a clustering process is an arduous task and there is a paucity in literature as enjoyed by the classifier algorithms. Previous works have shown that there is no single Cluster Validation Metric (CVM) that outshines the rest (Zhao et al., 2017). Nevertheless, it is important to outline CVM methods. There are three types of CVM, internal, external and relative validation. The internal CVM like Silhouette Coefficient, Dunn Index (DI), and Davies-Bouldin Index (DBI) rely upon the internal clustering information of the process without referencing any external information. This means that such indices use only information from the clustered datasets. They are usually based on minimization of inter-cluster distance, such as the Dunn index or on the average silhouette width as the silhouette index. Other methods related to external validation like Accuracy, Rand Index (RI), Adjusted Rand Index (ARI), Jaccard, Fowlkes-Mallows and Callinski-Harabaz Index (CHI) also known as variation of information criterion- evaluate a cluster division by a comparison with an already known correct partition or a known class variable akin to a classification task. The relative CVM evaluate the cluster by exercising varied parameter values in an algorithm (e.g., several reiterations of the cluster numbers). This thesis uses an internal CVM called the silhouette coefficient (Rousseeuw, 1987) also known as average silhouette width. It can be written as $\text{silhouette(k)} = \frac{1}{\text{n}}\sum_{\text{i=1}}^{\text{n}} \frac{\text{b(i)}-\text{a(i)}}{\max\,[\text{a(i)},\text{b(i)}]}$ where a_i is the average dissimilarity of the i-th object to the

other objects in the same cluster, b_i is the minimal average dissimilarity of the i-th object to any cluster not containing the i. The silhouette index takes value from -1 to +1. A value close to one indicates well-separated clusters; a value close to minus suggests badly separated clusters. A value close to zero indicates that the objects in the dataset are often located on the border of two natural clusters.

3.4. Chapter summary

This chapter presents the detailed description of each step of the proposed research approach. However, the next chapter presents a more detailed explanation of research contribution and the experimental models.

CHAPTER 4: ALTERNATIVE APPROACH TO UNSUPERVISED FEATURE SELECTION for MIXED DATA CLUSTERING

4.1. Introduction

An alternative approach to Unsupervised Feature Selection, based on the K-prototypes algorithm to improve the cluster purity of mixed dataset (UFSMDC), and reduce processing time is presented in this chapter. The summary of the contributions of this work are:

a. An alternative algorithm based on the K-prototypes method for mixed data clustering is proposed. It applies a modified Gower dissimilarity distance measure that can improve the computational complexity and algorithm run time.
b. The proposed algorithm is expected to perform better than the state-of-the-art partition-based clustering algorithms due to an integrated feature selection process.
c. The proposed algorithm is applied to real-world datasets (both educational and non-educational) for cluster purity.

4.2. The K-prototypes algorithm for mixed data clustering by Huang (1997)

The K-prototypes method was suggested by Huang (1997) to handle mixed dataset. It consists of the following steps; (1) Initially determine random cluster centres called the prototypes in a dataset X. (2) Then assign each data feature within X to a group whose antecedent is closes to it. The prototypes should be updated with each cycle of the

algorithm. (3) Once all the data features are assigned to respective clusters than re-compute the similarity between the data features with reference to the prototypes assigned in step 1. If a variable is inaccurately placed in a group, and that its nearest antecedent is associated to another group, then reallot the variable to that group and update the antecedents of both the groups. (4) Repeat step (3) till none of the variables have delineated from its group. The method involves three interconnected sub-steps, namely: the preliminary prototype identification, initial appropriation and re-appropriation. In the preliminary sub-step, a selection of K-random variables is committed as the initial group. Thereafter, this process is iteratively executed such that all the variables are segregated into their identifiable groups shown in Table 4.1.

Table 4.1: The K-prototypes algorithm by Huang (1997)

(1) Select K initial prototypes from a data set X, one for each cluster.
(2) Allocate each object in X to a cluster whose prototype is the nearest to it according to the given equation, $d(X_i, Q_1) = \sum_{j=1}^{m_r}(x_y^r - q_y^r) + y_i \sum_{j=1}^{m_c} \delta(x_y^c - q_y^c)$.Update the antecedent of the cluster after each allotment.
(3) Once all objects are allotted to a group than precompute the object similarity against the current antecedent. If an object is found such that its nearest antecedent belongs to another cluster rather than its current one, reallot the object to that cluster and update the antecedent of both clusters.
(4) Repeat (3) until no object has changed clusters after a full cycle test of X

4.2.1. The K-prototypes algorithm drawbacks

Since, this algorithm is based on the K-means algorithm it suffers from the problem of not knowing the initial number of groups. This has to be supplied by the user. Thus, it cannot guarantee a global optimum solution. The algorithm performance is O ((t+1) k*n), where n is the number of data points, K the number of groups and t is the number of

computational cycles of the re-electing process. Another problem with this method is the distance measure for categorical and numerical data. This algorithm applies the squared Euclidean distance to measure similarity between numerical data points and the conformity method for categorical features is the number of incorrect data points between variables and the cluster groups.

It is important to mention that Huang fail to discuss the validation of their results in the paper. The paper does not state the cluster evaluation criterion. Therefore, it is difficult to ascertain the accuracy or the purity of obtained clusters (Huang, 1997).

4.3. An improved K-prototypes clustering algorithm for mixed data clustering by Ji et al. (2013)

A group of researchers Ji et al., (2013) proposed an improvement to the K-prototypes algorithm in 2013. Their idea was to compute the fuzzy distribution centroid for nominal categorical features with the mean or average of the numerical features. A centroid is the geometric centre of an object. It was suggested by the authors to determine the frequency of categorical feature occurrence and then group them into distribution wise centroids. This way, almost all the characteristics of a categorical feature is captured, which, is then applied to determine the centre of a cluster. This approach is similar to the fuzzy centroid approach in (Kim, Lee, & Lee, 2004) by Kim et al. The Table 4.2 shows their proposed algorithm.

Table 4.2: An improved K-prototypes algorithm by Ji et al (2013)

Step 1. Suppose the agglomerative count of groups is k, and the maximum iterations is denoted as maxIte. Suppose, the initial number of clusters is λ, then we designate say k objects devoid of missing values and convert it to initial prototypes Q(t)=(Q1,Q2,…,Qk). This will generate random groups with initial significance values set to St={s1t,s2t,…,smt}(∑j=1msjt=1), and set t=0. Step 2. Fix Q', S' as Q^t and S^t respectively, minimize the problem $E(U,Q',S')$ to obtain U^{t+1}. Step 3. Fix U', S' as U^{t+1} and S^t respectively, minimize the problem E (U',Q,S') to obtain Q^{t+1}. Step 4. Fix U', Q' as U^{t+1} and Q^{t+1} respectively, minimize the problem E (U',Q',S) to obtain S^{t+1}. Step 5. If there is no improvement in E or maxIte equals to 0, then stop; otherwise, set t←t+1, maxIte←maxIte−1, and go to Step 2. The rule of conversion in Step 1 is described as follow: if the *j*th feature is the numeric in nature, then each $q_{ij} \in q_l$is the value of this feature; if the *j*th feature is the categorical in nature one, then each $c_{ij}^i \in Q_l$is assigned the value of 1.0 for $w_{lj}^k \; if \; x_{lj} = a_j^k; 0 \; for \; w_{lj}^k \; if \; x_{lj} \neq a_j^k$

4.3.1. Drawback of the improved K-prototypes algorithm

The drawback of this proposed approach is the suggested data conversion rule. This rule states that if the *j*th feature is numeric then then each qlj∈Ql is the value of this feature; if the *j*th feature is the categorical one, then each clj′∈Ql is assigned the value 1.0 for ωljκ if xlj=ajκ; 0 for ωljκ if xlj≠ajκ. This indicates that the approach is similar to the fuzzy centroid approach given by Kim et al (Kim, Lee, & Lee, 2004) . And because of this property, the complexity of algorithm raises to O(k(m+p+Nm−Np)nl), which is mainly attributed to the count of computational cycles 1 that are needed by the method to converge. The computational complexity of the K-prototypes method is O((l+1)kn) is high.

Regarding the cluster validation criterion, Ji et al (2013) used an external Cluster Validation Metric (CVM) called "accuracy". It is given as, $r = \frac{\sum_{i=1}^{n}=1a^i}{n}$ where a^i is the number of data objects occurring both in the ith cluster and its corresponding true class, and n is the number of data objects in the data set. According to this criterion, a higher value of r indicates a better clustering result with perfect clustering, if the value of$r = 1$.

Some questions were raised when using "accuracy" as a CVM for a clustering task. Foremost, it requires a class label to be known beforehand, which automatically defeats the purpose of clustering. Second of all, Ji et al referenced Huang & Ng (1999) that used accuracy as a CVM. The cited paper discussed a fuzzy K-modes algorithm for clustering categorical data, which is unrelated to the K-prototypes algorithm proposed by Huang in 1997. Moreover, Ji et al improved algorithm is based on the Huang's K-prototypes algorithm. But as given in sub-section 4.2.1, Huang did not discuss any CVM to test the validation of its results. To summarize, Ji et al based their approach on Huang's 1997 paper on K-prototypes but in the absence of a CVM in Huang's paper, they used accuracy as a CVM and as such tested their approach on a supervised CVM. Moreover, as discussed in section 2.8 the benchmark paper by Arbelaitz et al (2013) for CVM used in this thesis, also states that accuracy is a classification metric. Basis of these reasons, this thesis cannot compare the proposed method with Ji et al approach.

4.4. The Gower dissimilarity measure for mixed dataset by J.C. Gower (1971)

The original Gower coefficient (Gower, 1971) was introduced as a similarity measure. Let us assume a data set M consist of n numerical and p nominal features in a data matrix X given as $\boldsymbol{X} = [x_{ic}]$ where $i = 1,2,..,n$ (n is the total number of objects) and c= $1,2,..,m$ (m is the total number of variables). Then similarity between objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are characterized by values of mixed data variables, is expressed by the formula

$$S_G(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\sum_{c=1}^{m} w_{ijc} s_{ijc}}{\sum_{c=1}^{m} w_{ijc}} \quad (1)$$

where s_{ijc} is a similarity measure between the i-th and j-th objects by the c-th variable$c = (1,2,..,m)$, and w_{ijc} takes the value zero, if either the i-th and the j-th objects by the c-th variables is missing; otherwise it takes the value one.

If the c-th variable is nominal then similarity between its two categories is treated as zero for matches of categories, and as one otherwise. If the c-th variable is numeric, similarity is expressed by the formula

$$s_{ijc} = \frac{|x_{ic} - x_{jc}|}{\max(x_c) - \min(x_c)} \quad (2)$$

If the c-th variable is ordinal, all categories are transformed according to the formula

$$x_{ic} = \frac{r_{ic} - 1}{R_c - 1} \quad (3)$$

where r_{ic} is the rank number of the ith ordinal category$(r = 1,2, \dots, R_c)$, and the R_c is the minimal rank number of the c-th variable. Basis of this transformation, the outcome values in Equation (2) can be used for numeric variables.

The idea was to calculate the similarity between values contained in each variable and then average them across all the variables. When all the variables are quantitative in nature, then the coefficient is range-normalized by applying the Manhattan distance. The coefficient range is between 0 and 1.

4.4.1. Drawback of the Gower Dissimilarity Measure

The Gower dissimilarity is simply 1-Gower Similarity (GS). This means the limitations of GS are the same for Gower dissimilarity. The range normalization of quantitative variables causes information loss. By *origin*, Gower dissimilarity is non-Euclidean and non-metric (even when all variables to compute it had been interval, Gower index will be closer to Manhattan distance, not Euclidean distance). The Gower distance, without

ordinal variables present (i.e. w/o using the Podani's option) $\sqrt{1-GS}$ behaves as Euclidean distance, it fully supports Euclidean space. But $1-GS$ is only metric (supports triangular inequality), not Euclidean. With ordinal variables present (using the Podani's option) $\sqrt{1-GS}$ is only metric, not Euclidean; and $1-GS$ isn't metric at all. With Euclidean distances (distances supporting Euclidean space), any classic clustering method will work, including K-means (if K-means can process distance matrices, of course). Using K-means or other those methods based on Euclidean distance with non-Euclidean still metric distance is heuristically admissible, perhaps. With non-metric distances, no such methods may be used. This formula will only work for datasets with complete entries.

4.5. The Proposed Unsupervised Feature Selection Algorithm for Mixed Data Clustering (UFSMDC)

On the basis of the K-prototypes algorithm and the modified Gower coefficient, an improved partition-based feature selection algorithm is proposed. This section begins with a brief introduction of some preliminaries in section 4.5.1. The section 4.5.2 shows the flowchart of the proposed approach. In section 4.5.3, the proposed approach is detailed. The section 4.5.4 discusses the difference between the proposed approach and the algorithm by Huang (1997). The comparison result is discussed in section 4.5.5. Finally, the performance analysis of the proposed algorithm is discussed in sub-section 4.5.6.

4.5.1. Preliminaries

Suppose a dataset D consist of I instances. Each instance with n features (n_{cat} categorical and n_{con} numerical) where $D_n(1 \leq n \leq I)$ denotes the *n-th* feature. The numerical features are standardized to a median scale. For simplicity, the categorical features are set before the numerical features.

Definition 1. For a cluster C and n feature value $n_i \varepsilon D_i$, the frequency of n in C with respect to D_i is defined as: $Freq_{C|D}(\mathrm{n}_i) = |\{\text{instance} \mid \text{instance} \in \mathrm{C}, \text{instance}.\mathrm{D}_i = n_i\}|$

Definition 2. For a cluster C, the cluster gist (CG) is defined as: $CG = \{\mathrm{p}, \text{gist}\}$ where p is the size of the cluster $C(\mathrm{p} = |C|)$, gist is given as the frequency of information for categorical features and centroid for numerical features:

$$dif(\mathrm{C}_i^{(1)}, \mathrm{C}_i^{(2)}) = 1 - \frac{1}{|C_1|.|C_2|} \sum_{p \in C_1} Freq_{C_1|D_i}(p_i).Freq_{C_2|D_i}(p_i) = 1 - \frac{1}{|C_1|.|C_2|} \sum_{r \in C_2} Freq_{C_1|D_i}(\mathrm{r}_i).Freq_{C_2|D_i}(\mathrm{r}_i)$$

The modified Gower distance approach detailed in sub-section 4.5.3, is applied as a dissimilarity measure in this thesis because it can work for both categorical and numerical features. The distance between n_{cat} and n_{cont} is given by $d(i,j) = \frac{\sum_k \delta_{ijk} d_{ijk}}{\sum_k \delta_{ijk}}$. In particular the d_{ijk} represent i^{th} and j^{th} unit computed considering the k^{th} variable. This dissimilarity measure can be used to determine how dissimilar two different observations are. The observation may contain combination of logical, numerical or categorical data. The distance is always a number between 0 (identical) and 1(maximally dissimilar).

4.5.2. Flow chart of the proposed approach

Based on the definitions 1, 2 and the K-prototypes method discussed earlier, an improved partition-based feature selection algorithm is formulated. The flowchart of the proposed approach is given in Figure 4.1.

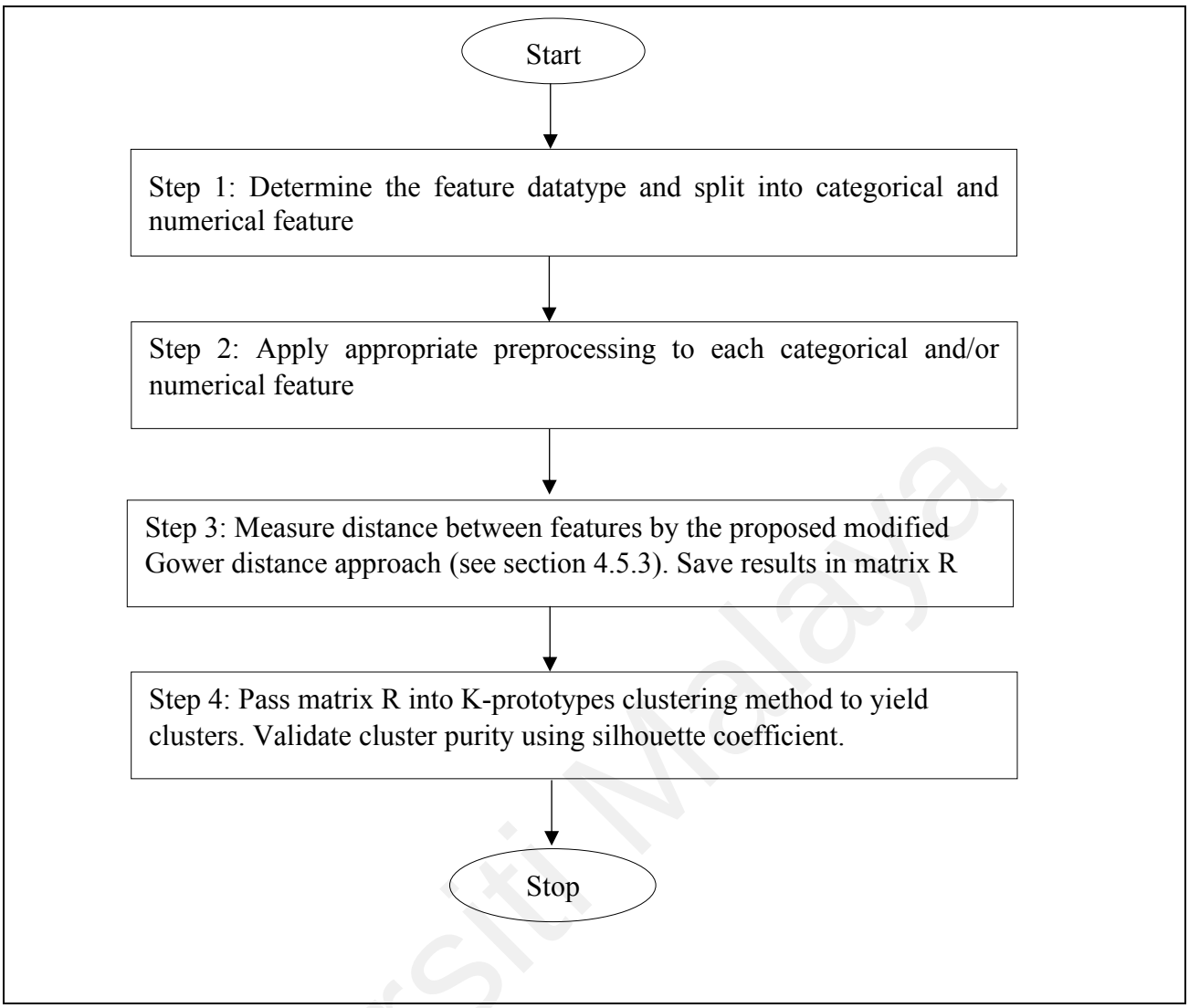

Figure 4.1: Flowchart of the proposed approach

Step 1: Calculating distance

After the initial pre-processing, the distance between the variables is calculated (numerical and categorical) by applying the proposed modified Gower distance elaborated in sub-section 4.5.3. The original Gower distance is calculated as the average of partial dissimilarities across individuals. In a partial dissimilarity, a particular standardisation is applied to each feature, and the distance between n features is the average of all feature-specific distances. For example, a numerical feature C_n and C_{n1} partial dissimilarity is the ratio between 1 minus the absolute difference of observations C_{na} and C_{nb} where n being the number of observations in the dataset. In this work, the numerical feature

is standardized on median. The distance between numerical features n_1 and n_2 is calculated by Manhattan distance. The nominal features are normalized on frequency centroid. Later using equation (4) detailed in sub-section 4.5.3, the frequency centroid of nominal features was calculated. It is an iterative process that repeats n number of times using a strategy that builds on the principle of leave-one-out feature elimination over W in measuring the importance of each feature f_p.

Step 2: Choosing a clustering algorithm

The reduced feature set f_p is passed into the K-prototypes method (Huang, 1997), which has been discussed on page 15, sub-section 2.4.1.

Step 3: Selecting the number of clusters

This thesis determines the possible number of groups or clusters by using the Elbow method (Kaufman & Rousseeuw, 2009).

4.5.3. The proposed approach

Suppose a data set M consist of n numerical and p nominal features in a matrix X given as $M = \{\mathrm{n}, p\}$. The numerical features are normalized on median and the distance between numerical features n_1 and n_2 is calculated by Manhattan distance. The nominal features are normalized on frequency centroid. A frequency centroid is written as:

$$f_c = [fp_1, fp_2, ..., fp_n] \tag{2}$$

Where f_c is the cumulative frequencies and fp_n is the frequency of occurrence for the p_n feature. Then substituting the parameter s_{ijk} in equation (1) with equation 2, the modified equation is:

$$S_{ij} = \sum_{i=1}^{n} fp_{ijk}\delta_{ijk} \Big/ \sum_{i=1}^{n} \delta_{ijk} \tag{3}$$

Using (3), this thesis proposes a quantitative measurement to determine the importance of each feature fp_{ijk}. In doing so there can be two cases:

- Case 1: $fp_{ij} > 0$, in this case Fp_{ij} is a high variance feature
- Case 2: $fp_{ij} < 0$, in this case Fp_{ij} is a low variance feature

The idea is to apply association to determine features with variance and to ensure the inherent correlations are preserved.

The modification of Gower similarity coefficient is made to the nominal part of the coefficient. It assigns lower weights to the matches in variables with high variance, because they occur in high frequency and overshadow the lower occurring (low frequency) variables. It can be expressed by the formula given in equation (4):

$$s_{ijc} = \left\langle \begin{matrix} 0 & if x_{ic}=x_{jc}, \\ 1-\frac{1}{1+lnf\left(x_{ic}.lnf\left(x_{jc}\right)\right)} & \text{otherwise} \end{matrix} \right\rangle \tag{4}$$

where $f(x_{ic})$ is an absolute frequency of the value x_{ic} in the c-th variable. The similarity measure takes the value zero in case of match of categories, and the values from zero to number $1 - 1/(1 + \ln{(n/3)^2})$ otherwise, until the dataset size converges to one. The Figure 4.2 presents the mechanism of the proposed approach.

Algorithm: Unsupervised Feature Selection for Mixed Data Clustering (UFSMDC) Input: $X: m.n$ dataset with m objects and n features $D\{F_1, F_2, \dots, F_n\}$ //a training dataset n is the number of clusters determined by Elbow method Output: n significant features in disjoint cluster(s)
Given a mixed dataset D_{mix} as input, 1. Pre-processing step: a. Determine the datatype of features and split into categorical ca_i and numerical co_i b. For each ca_i check and resolve issues like missing values, effect size or the strength of association and correlation significance. Save result in matrix D_{ca} c. For each co_i check and resolve issues like skewness, kurtosis, multicollinearity, outliers and missing values. Save result in matrix D_{co} d. Measure association between D_{ca} and D_{co} with factor analysis. Save result in M 2. Distance measurement step: a. Pass the matrix M to the modified Gower equation in (4) (*see section 4.5.3, equation 4*) b. Save result in matrix R 3. Clustering step: a. Pass matrix R and n into K-prototypes to yield clusters b. Validate cluster purity using silhouette coefficient End algorithm

Figure 4.2: An unsupervised feature selection approach for mixed data clustering

The idea is to preserve the association between categorical features with variance by maintaining the inherent correlations between the features. The proposed approach considers both complete and incomplete datasets. A complete dataset is devoid of missing values and an incomplete dataset consist of missing values.

4.5.4. The difference between proposed approach and Huang's approach (1997)

The proposed approach differs from the algorithm proposed by Huang (1997) in the following ways:

a. The proposed approach considers the feature properties and measured the association between them which reduces the probability of highly correlated features. As such it clearly indicated that the proposed approach reduces the algorithmic computation as compared to Huang (1997).
b. The proposed approach also considers the outliers in the feature set and treats them instead of removing them. Subsequently, this improves the cluster cohesiveness by retaining as much viable information contained in the features. Hence it reduces time and improves the cluster purity.
c. The approach of Huang (1997) are focused on the data clustering and not on feature selection.

4.5.5. Comparison of the results

The proposed approach is compared with the algorithm by Huang (1997). The computational complexity of the algorithm of Huang (1997) has the following basic operation:

a. Initial prototype (or the number of features) selection
b. Initial allocation of features to *n* clusters
c. Re-allocation of features by finding the object similarity measure for both categorical and numerical features.

The space time complexity of this method is O((t+1)kn), where the count of data points is *n*, *k* is the count of groups and *t* is the count of cycles in the reallocation process. Ideally, k << n and t seldom exceed more than a hundred iterations.

The algorithmic complexity of creating new groups and revising them is O(kmn) and O(k(p+Nm−Np)n). Where k is group enumeration; p is the enumeration of numeric data points; m is the enumeration of all data points; N=max(t) is the highest frequency of categorical feature values; and n is the enumeration of all data points. So, we can say the overall time complexity is O(k(m+p+Nm−Np)nl), where l is the iteration tally responsible for the method to converge. The computational complexity of this method is O((l+1)kn). And as such its computational complexity is higher than the K-prototypes. The space complexity of this algorithm is O(k(p+n+mN−pN)+mn). The algorithm follows a minimum of three basic operations as outlined above, but both of them do not consider the feature properties of a feature. Based on these observations, it can be deduced that the initial number of cluster selection needs to be provided by the user (*and is thus random in nature)*. Moreover, the number of features to be accessed once in every iteration is *m.n* and comparing this with the proposed algorithm in which no such computation is required. Thus, the amount of computation is significantly reduced. The number of iterations performed by each algorithm in Figure 4.3, shows the complexity ratio, as the number of features increase, the number of iterations in K-prototypes increase, while in our proposed method is significantly reduced.

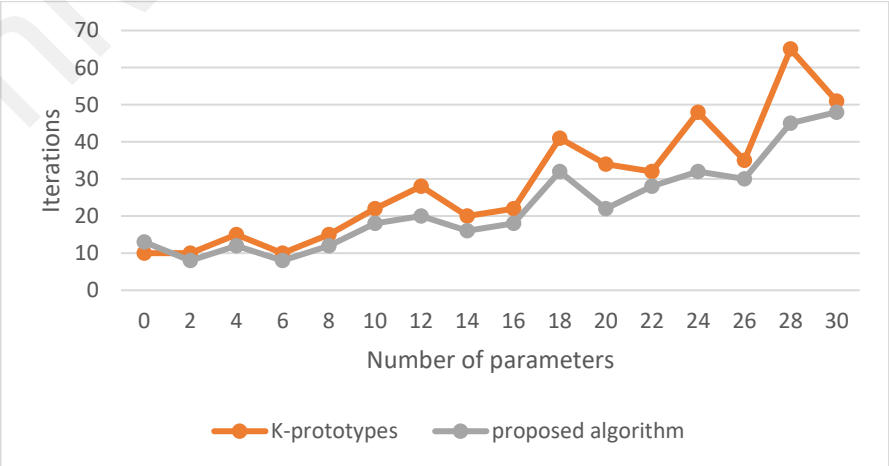

Figure 4.3: Number of iterations performed by the two algorithms

4.5.6. Performance analysis of the proposed approach

Feature selection is essentially a combinatorial optimization problem. The objective is to reduce the number of features such that maximum variance amongst the features are preserved while accounting for aspects such as speed, efficiency and computational speed. Several algorithms exist to handle the reduction of features, but it seems they did not solve the issues of optimality in which every dataset could be reduced.

The computational complexity of K-prototypes is determined by counting the number of basic operations. The basic operation in this case is simply the number of times each feature is accessed. For UFSMDC algorithm, the only features with high degree of association are considered and are accessed only once. And then finding the subset $Y = \{a'_1, a'_2, \dots, a'_n\} \in E\ for\ 1 \leq i \leq m$ is also accessed once.

Therefore, $UFSMDC = m.n + n$

Suppose $m = n^2$ then $n^2 + n$. Hence the complexity of UFSMDC is $O(n^2)$. And therefore, comparing the complexity of O ((t+1) kn) and O (k (m+p+Nm−Np)nl), the UFSMDC decreases the computational complexity. The Table 4.3 details the comparative summary of the three algorithms which is dependent on the iteration count, computational complexity and feature selection.

Table 4.3: The comparison results

Comparison	K-prototypes	UFSMDC	Remark
The operation involved	• Random feature selection. • Random cluster allocation	• Feature selection based on feature properties and association strength • Outliers and missing values are treated and not removed	Both algorithms require some certain set of operations.
Computational complexity	O((t+1)kn)	$O(n^2)$	The K-prototypes and UFSMDC have lower computational complexity
Limitation	Remove features with missing values causing information loss	Dependent on distribution of data and factor levels for categorical data	The priority of K-prototypes does not maintain a complete feature-set.
Feature selection result	Information loss is evident as only complete feature set are retained	Information loss is minimized by retaining incomplete feature set including the complete feature set	UFSMDC maintain optimal variant feature set while K-prototype does not.

4.6. Summary of the chapter

This chapter explains the voluminous work in mixed data clustering with a focus on educational data mining context. It has presented an algorithm that overcomes the problems of the existing feature selection method for mixed data clustering. It shows a comparison between the proposed approach with existing partition clustering methods that can handle mixed data ranging from computational complexity, difficulty in understanding of the algorithms and implementing it within any dataset with ease. It also explains the difference between the proposed approach and the baseline methods that were used for comparison.

CHAPTER 5: APPLICATION OF UFSMDC APPROACH ON MIXED DATASETS AND RESULT

5.1. Introduction

This chapter details the different experimental setup carried out for the implementation and validation of the proposed approach. The dataset used is described. The baseline methods are disclosed, and the performance evaluation metrics are explained. It also discusses the different experimental results of the study.

5.2. Baseline Methods

This research has reported three (3) different experiments, each incorporating a different set of feature type in the clustering process. In assessing the effectiveness and efficiency of these experiments, besides comparing each experiment with one another, it also compares with two other baseline methods presented in Huang (1997) and Gower (1971) that are discussed in section 2.4.3

5.3. Performance Evaluation Metrics

This thesis applies an internal CVM, the Silhouette Coefficient (SC). The reason SC is used because the class label or the target class is unknown in advance. A detailed justification on the choice of SC is given in chapter 2, sub-section 2.8. The best value of SC is 1 and the worst value is -1. The SC values near zero indicates overlapping clusters. The SC values near 1 indicates pure clusters. Where purity is defined by the similar objects close to each other within the cluster. A detailed discussion on existing CVM is present in section 2.3.

5.4. Application of the proposed approach on mixed educational dataset

A typical student record consists of student's address (*categorical features)* and examination results *(numerical features)*. Data that consist of both numerical and categorical features is called a mixed dataset. Much of the existing research in EDM is focused around univariate or multivariate data. However, in this experiment a mixed educational dataset is used. The sub-section 2.4 until 2.8.3 presents a brief discussion on multivariate and mixed dataset. The problem statement for this experiment is stated as follows:

Given a multivariate educational dataset consisting of both numerical and categorical data types, retrieve the features accounting for maximum variance between each other. For each of the numerical feature resolve issues like skewness, kurtosis, multicollinearity, outliers and missing values. For each of the categorical features, resolve issues like effect size, strength of association, high correlation and missing values. Determine the strength of association amongst the numerical and categorical feature set and recommend the top-N most disjoint clusters.

5.4.1. Experimental Setup

This research has utilized a school panel level dataset for academic session 2012-2013 for the state of Delhi, acquired from the District Information System for Education (DISE) (Azam & Saing, 2017). The dataset consisted of six comma separated data files on school demographics (such as school name, location, and address), school facilities, enrolment and repeater records as well as teacher data. Incidentally, these six files contained different types of data for the same school, because, they all had a common 10-digit school code. These six data files were merged into a single data file on basis of common column. This resulted in 183 features for 5103 schools. Of these 5103 schools there were 2581 primary level (grade 1-grade 5), 518 primary to upper-primary (grade 1-grade 8), 1014 primary to higher secondary (grade 1- grade 12), 39 upper-primary to higher secondary level (grade 6-grade 12), 504 upper-primary level only (grade 6-grade 8). For this work, the focus is on primary level schools because the interest is in analysing and comparing the factors

responsible for student enrolment in primary level schools. From 963 pre-primary level (or kindergarten) schools in the 2581 primary level schools, this work removed the kindergarten schools to obtain 1618 primary level schools in 183 variables. The dataset consisted of features with both categorical and numerical data types. It needs to be mentioned that the categorical features were of nominal data type. A nominal categorical feature does not have any intrinsic order to it. Example of a categorical feature is colour which can have entries such as red, blue, green and yellow. There is no order to it, hence it is a nominal feature. For validation purpose, this work has deposited the clean data files on IEEEDataPort[1]. Table 5.1 depicts the statistical properties of the dataset used in this work.

Table 5.1: Statistical properties of the dataset

Total count of observations	1469
Total count of variables	133
Total count of categorical variables	17
Total count of numerical variables	116
Total count of variables with zero variance	84
Total count of variables with variance	49
Total count of missing values	0

[1] http://ieee-dataport.org/1219

The Figure 5.1 and 5.2 shows the primary schools demographic and enrolment distribution.

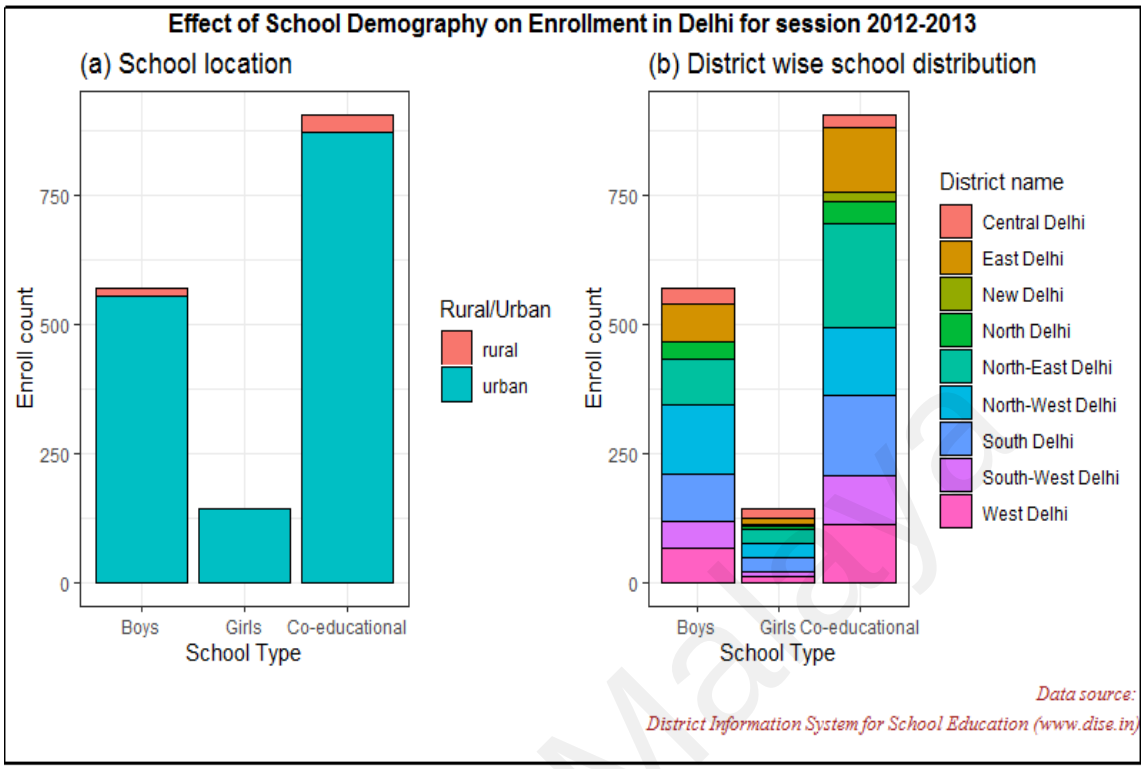

Figure 5.1: Primary school enrolment and demographic distribution

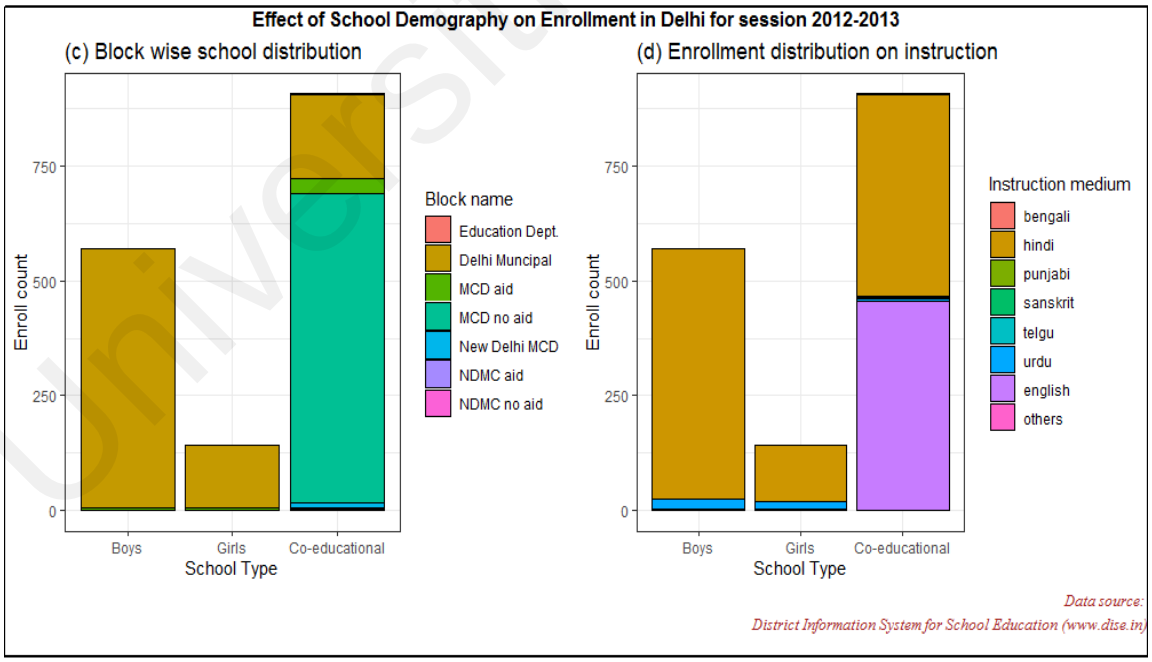

Figure 5.2: Primary school enrolment and school facilities distribution

This experiment starts by determining the data type of the feature and splits them accordingly into two separate data frames. The numerical features are normalized on median and the distance

between numerical features m_1 and m_2 is calculated by Manhattan distance. The nominal features are normalized on frequency centroid. Both feature types are treated for issues like collinearity, multicollinearity, skewness and missing value imputation. Thereafter, both the categorical and numerical data frames are merged and passed into the modified Gower distance method proposed in equation (3) and save the result in a data matrix. This data matrix is then passed into the K-prototypes algorithm to yield clusters. The obtained cluster purity is validated using SC. Finally, the system presents the clusters to the user. To illustrate the approach further, Figure 5.3 and Figure 5.4 represents the outliers present in numerical features in the primary school panel level dataset. In the Indian education system, the primary school comprises of class 1 till class 5.

To determine the outliers for numerical variable, it is often noted in literature to use Interquartile range as a metric, where outlier values are those that lie outside of $1.5 * IQR$. The points outside the whiskers in the box plot shown in Figure 5.3 and Figure 5.4, denoted as dots are the outliers. The goal is to determine the features that are not redundant and account for maximum variance. The rationale behind this approach is features that are redundant or are highly collinear to each other only add noise to the resultant model. Therefore, an unrelated feature-set makes a pure cluster. Applying this approach, the data dimensionality is reduced, as shown in Figure 5.5 and Figure 5.6 and also retain numerical features accounting for maximum variance amongst them.

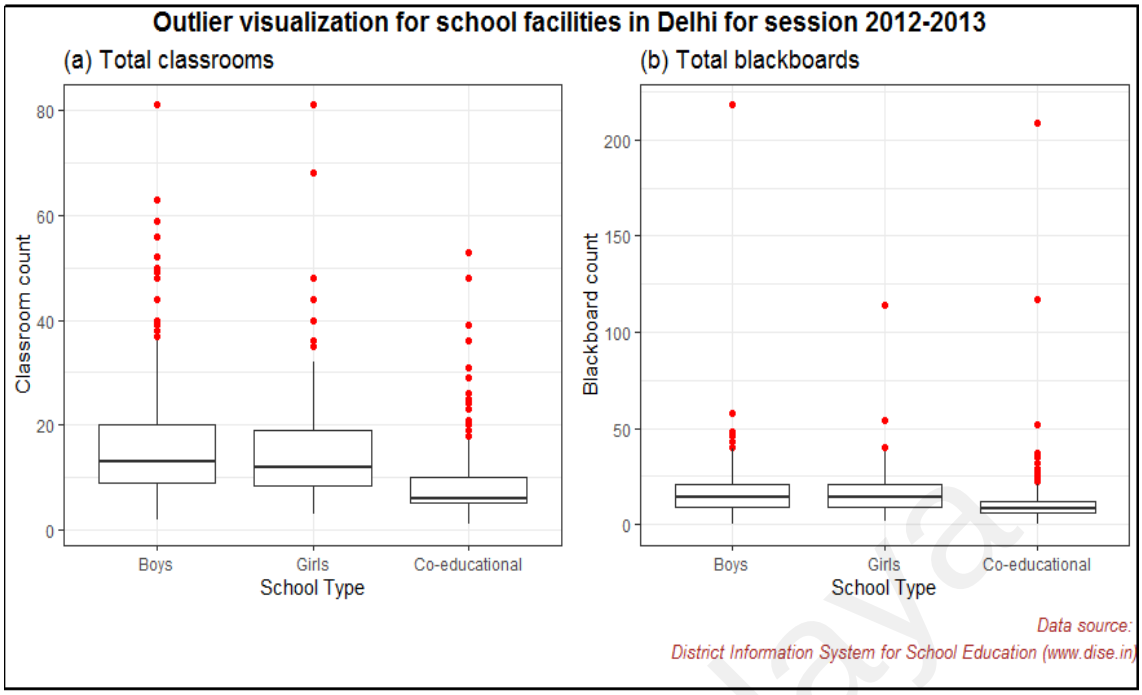

Figure 5.3: Outliers in numerical features

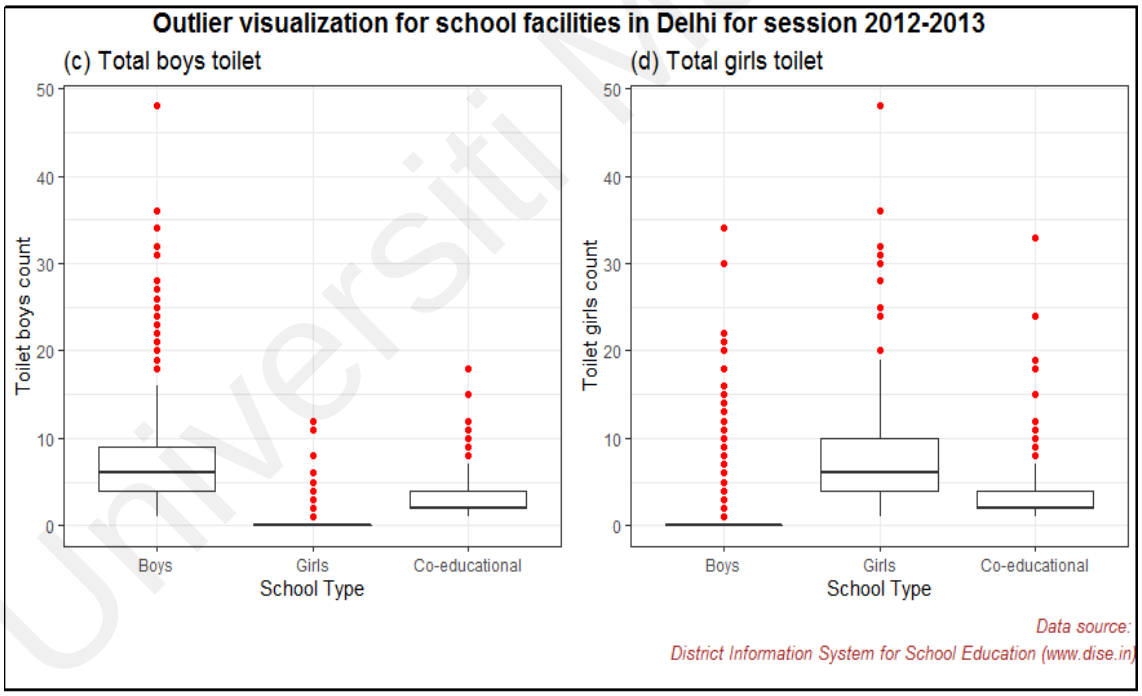

Figure 5.4: Outliers in numerical features

A partial outlier treatment was performed depicted in Figure 5.5 and Figure 5.6; the data dimension reduced to 1,237 primary schools in 51 features. Of these, there were 401 boys' schools, 90 girls' schools, and 874 co-educational schools. The boxplot in Figure 5.3 (a), shows the median for total number of classrooms in boys' school is 16 and 15 for girls' school. When the outliers are partially

treated for, shown in boxplot in Figure 5.5 (a), a median of 10 classrooms per boys' and girls' school is obtained. Therefore, by following this approach this work was able to balance the distribution of data such that it captures the maximum variance.

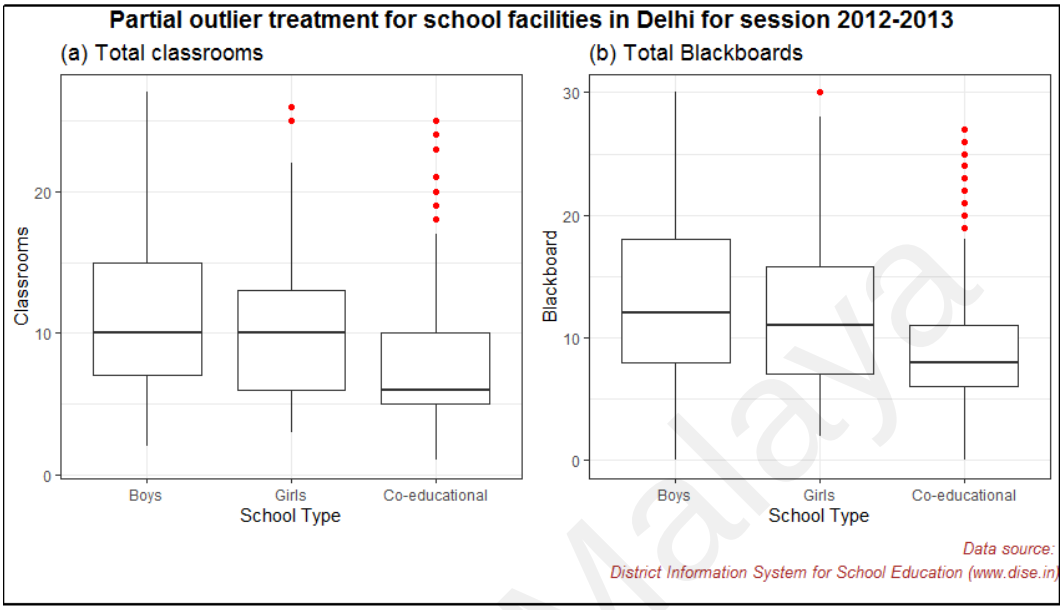

Figure 5.5: Partial outlier treatment for numerical features

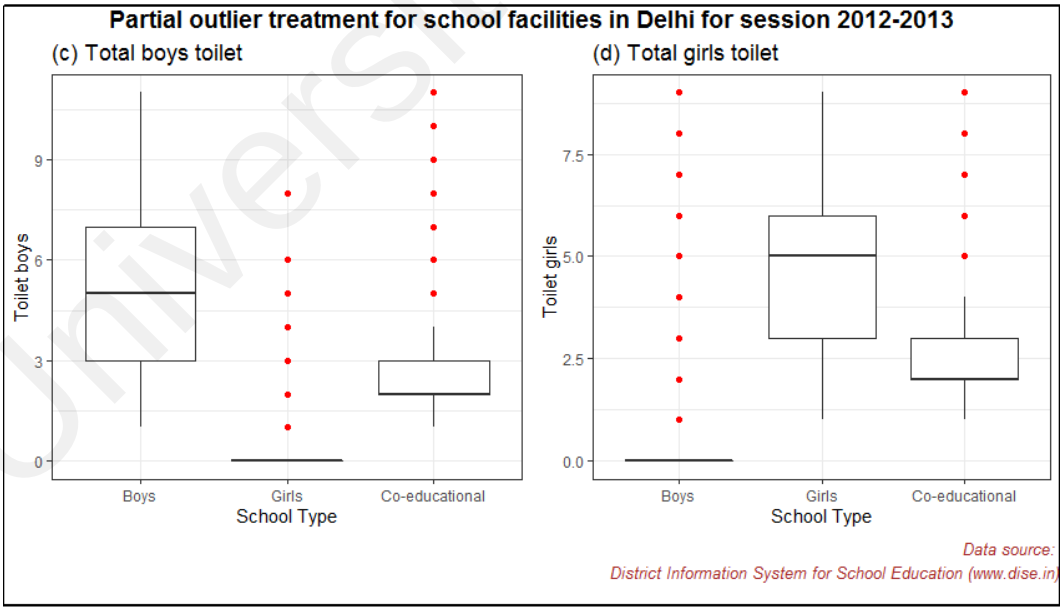

Figure 5.6: Partial outlier treatment for numerical features

The need to group multiple variables of different data types was imminent as this involved a mixed dataset. Therefore, clustering allowed using multiple features to identify similar groups in an

unsupervised fashion. The proposed approach has utilized the underlying association to determine features with variance and to ensure the inherent correlations are preserved. Such features will then be passed into the modified Gower distance function given in equation (3) to yield the similarity matrix W from the dataset D. From W, a normalized matrix was developed for numerical features. Later using equation(2), the frequency centroid of nominal features was calculated. It is an iterative process that repeats n number of times using a strategy, which builds on the principle of leave-one-out feature elimination over W in measuring the importance of each feature f_p through equation (3). Finally, the reduced set of features in f_c are assigned ranks from the most to the least importance based on the values obtained by f_p The reduced feature set f_p is then subjected to an internal cluster validation metric, the SC to determine the clustering purity. It also measures how similar an observation is to its own cluster compared its closest neighbouring cluster. The metric can range from -1 to 1, where higher values are better. After calculating SC for clusters ranging from 2 to 10 for the K-prototypes algorithm, it can be observed from Figure 5.7, that 3 clusters (government-owned schools, semi-government schools and private schools) yielded the highest variability.

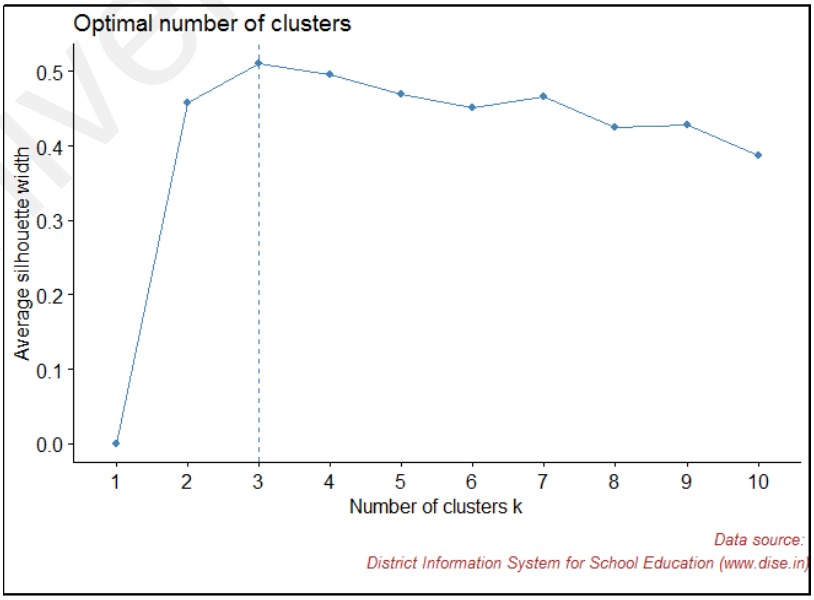

Figure 5.7: Determining the number of clusters using silhouette width

One advantage of the proposed approach is its ability to reduce the information loss sustained in data type conversion. The next experiment adds stricter rules in measuring the association strength between the numerical and categorical features to improve heterogeneous cluster purity.

5.4.2. A brief description of other mixed datasets

This thesis has used 04 mixed datasets from the UCI ML repository to evaluate the performance of the proposed algorithm. Their characteristics are described below.

a) Automobile: Is a mixed dataset. The data concerns city-cycle fuel consumption in miles per gallon. It consists of 8 features of which 3 are multivalued discrete and 5 are numerical features. The number of observations is 398 and it has missing values.
b) Auto mpg: Is a mixed dataset. The data concerns automobile characteristics. It consists of 26 features of which 11 are categorical and 15 are numerical features. The number of observations is 205 and it has missing values.
c) Census income: Is a mixed dataset. The data concerns household income details. It consists of 14 features of which 8 are categorical and 6 are numerical features. The number of observations is 48842 and it has missing values.
d) Credit approval: Is a mixed dataset. The data concerns credit card applications. It consists of 15 features of which 6 are categorical and 9 are numerical features. The number of observations is 690 and it has missing values.

5.4.3. Application of the proposed approach on other mixed datasets

This section discusses the experimental results of the proposed UFSMDC approach presented in section 4.5.1. To further evaluate the proposed approach, this work selected 04 mixed datasets (datasets with categorical and numerical features) from UCI Machine Learning repository. The sub-section 5.4.2 details their properties. Continuing

further, in literature there exist several partition-based clustering algorithms namely, K-means, K-modes, Partition Around Medoids (PAM), fuzzy c-means and Clustering LARge Applications (CLARA). Since K-means and Fuzzy C-means method worked for numerical data only, they were eliminated from the comparison. The K-modes algorithm worked only for categorical data. So, it was also eliminated from the comparison. The PAM, CLARA and K-prototypes can work for mixed datasets. A fixed number of four clusters is used in these experiments for fair comparison. In Table 5.2, the application of PAM and CLARA on 04 UCI ML mixed datasets and the school panel level dataset is given.

Table 5.2: Application of PAM, CLARA with and without proposed UFSMDC approach on mixed data

S. No.	Dataset name	Algorithm	Cluster Validation Metric-SC (without UFSMDC approach)	Cluster Validation Metric-SC (with UFSMDC approach)
1.	Automobile	PAM	0.55	**0.84**
		CLARA	0.57	**0.91**
2.	Auto mpg	PAM	0.53	**0.54**
		CLARA	0.55	**0.6**
3.	Census income	PAM	**0.93**	0.53
		CLARA	**0.87**	0.54
4.	Credit approval	PAM	**0.74**	0.32
		CLARA	**0.77**	0.35
5.	DISE School panel level dataset	PAM	0.35	**0.62**
		CLARA	0.38	**0.69**

From Table 5.2, it can be seen that most mixed datasets have obtained better results. However, the credit approval and the census income datasets, when subjected to the proposed approach, do not reveal any improvement. The reason is both these datasets have some categorical variables whose factor levels are in a very high number. Consider an example; student education level is a categorical variable with (factor) levels like kindergarten, primary, upper primary, secondary, higher secondary,

secondary, bachelor, master, PhD, post doc, some college, professional qualifications like diploma, certificate etc. are indicators of different educational levels. The focus here is on the several factor levels. Both these datasets including the school panel level dataset have a huge number of factor levels for most categorical variables. The proposed approach is able to process a categorical variable with utmost 10 factor levels, beyond that it will work but will affect the cluster purity. And this is the reason that why for these two datasets, it has recorded a decreased cluster purity. For the sake of additional discussion, it's worthwhile to state, partition clustering algorithms like PAM, CLARA and K-prototypes play no role in discerning the data properties be it categorical or numerical. They only divide the data into groups. Such groups are difficult to interpret in the absence of data properties like factor levels being considered.

5.5. Results Comparison and Discussion

This section discusses the application of the K-prototypes algorithm Huang (1997) using the Gower coefficient on the school panel level dataset. It then discusses the comparative results by integrating the proposed UFSMDC approach in the K-prototype algorithm for the same dataset.

Initially, the Gower approach was applied to calculate the distance between mixed variables. Using the Elbow method shown in Figure 5.7, three clusters were chosen. Thereafter, the K-prototypes algorithm was applied to yield clusters. These clusters were tested for purity using SC. It was found that results kept changing on each algorithm run cycle. Therefore, for result reproducibility a constant seed value was incorporated. The experiment was run 5 times with a different seed value on each run yielding a different SC value each time, which was recorded in Table 5.3.

Table 5.3: Five times execution cycle of Gower coefficient in K-prototypes on school panel level dataset

Algorithm execution cycle	Seed value	SC using Gower coefficient in K-prototypes
1	101	0.38
2	201	0.43
3	301	0.40
4	401	0.43
5	501	0.41

The average of the five SC values given in Table 5.3 was calculated and is shown in Table 5.4.

Table 5.4: Application of K-prototypes algorithm using the Gower coefficient on school panel level dataset

clusters	**No pre-processing**		**Algorithm using Gower coefficient**	**Cluster Validation Metric**
	Number of categorical features	Number of numerical features	K-prototypes	SC (average value)
3	17	116		0.41

Thereafter, the proposed approach as discussed in sub-section 4.3 was applied to calculate the distance between mixed variables in the school dataset. Again, using the Elbow method shown in Figure 5.7, three clusters were chosen. The K-prototypes algorithm was applied to yield clusters. The modified Gower coefficient method was executed five times with a different seed value on each algorithm run cycle. It yielded a different SC value and is shown in Table 5.5.

Table 5.5: Five times execution cycle of modified Gower coefficient in K-prototypes on school panel level dataset

Algorithm execution cycle	Seed value	SC using modified Gower coefficient in K-prototypes
1	101	0.47
2	201	0.44
3	301	0.46
4	401	0.43
5	501	0.44

The average of these SC values was computed to be 0.448 and rounding it off to two significant digits, it becomes 0.45, as given in Table 5.6.

Table 5.6: Application of K-prototypes algorithm using the UFSMDC approach on school panel level dataset

clusters	**After pre-processing**		**Algorithm using UFSMDC approach**	**Cluster Validation Metric**
	Number of categorical features	Number of numerical features	K-prototypes	SC (average value)
3	17	32		**0.45**

The original Gower distance is calculated as the average of partial dissimilarities across individuals. In a partial dissimilarity, a particular standardization is applied to each feature, and the distance between *n* features is the average of all feature-specific distances. For example, a numerical feature C_n and C_{n1} partial dissimilarity is the ratio between 1 minus the absolute difference of observations C_{na} and C_{nb} where n being the number of observations in the dataset. In the proposed approach, the numerical feature is standardized on median. The distance between numerical features n_1 and n_2 is calculated by Manhattan distance. The nominal features are normalized on frequency centroid. Moreover, the Gower approach does not consider the nominal categorical feature in a mixed dataset.

Such features were transformed into continuous format, causing information loss. Continuing further, the importance of seed value is actually relevant for the partition clustering algorithm and not the distance function. This thesis is focussed on the distance function for mixed data. The results obtained by the proposed distance approach for mixed dataset are better than the Gower distance is because of the modification to the nominal part in the Gower similarity coefficient. As stated in section 4.5.3, the equation 2, 3 is used to derive equation 4. Essentially lower weights are assigned to high frequency nominal features and high weight to nominal features with low frequency. Assigning weights balances the nominal feature frequency distribution across a sample space. It also helps in preserving the association between nominal features with variance by measuring the association strength between them. The Gower coefficient does not consider nominal feature, nor does it consider association strength and another drawback is the quantitative variables are range normalized. Whereas, in the proposed approach they are normalized on the median. Thereafter, as shown in equation 4, the similarity between nominal features S takes the value 0 case of match of categories, and the values from zero to number $1 - 1/(1 + \ln (n/3)^2)$ otherwise, until the dataset size converges to one.

The improved K-prototypes algorithm by Ji et al (2013) is supervised classification and not a pure clustering algorithm. The detailed explanation is given in chapter 4 sub-section 4.3.1, where this work has explained the drawbacks of their approach. While they applied an external CVM to evaluate their method, this thesis used an internal CVM. They used four datasets (iris, heart disease, soybean and credit-approval) from UCI ML repository. Off these, the iris dataset is purely numerical in nature. The Soybean dataset is purely categorical and the remaining two datasets, heart disease and credit approval are mixed data. To calculate the distance between categorical variables, they have used fuzzy centroid which they call hard clustering, but, fuzzy in itself is a soft clustering approach. For these reasons, this thesis does not compare the proposed approach with the Ji et al (2013) approach.

This research suggests some practical guidelines for educational data miners or users of educational datasets such as:

a. The educational data miners can focus on building semantic educational recommender system-SERS which is in tandem with the service-oriented approach of the third generation learning management systems, where external educational web-based services can interoperate with the learning management systems.
b. The e-learning systems can integrate information system modules that focus on personalising content delivery based on the learner's interaction with the system. They can also include automated personalised feedback mechanisms to provide feedback to learners at scale. These systems should be designed such that they can capture both the learners implicit and explicit interactions with the e-learning system.
c. The traditional classroom-based learning environment can make use of video recording system to record classroom activities. Such recordings can later be studied by both the teacher and the researcher to learn and improve teacher-student interactions.

5.6. Chapter Summary

This chapter discusses the experimental results of UFSMDC approach in a real -world educational dataset as well as several benchmarked mixed datasets. The proposed approach was integrated in an existing partition clustering algorithm, the K-prototypes. A comparative evaluation of experimental results in applying only the K-prototypes algorithm on mixed datasets and then integrating the proposed approach as a precursory step before applying the K-prototypes algorithm, reveal the superiority of the approach.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1. Introduction

This chapter concludes the research on a partition-based feature selection approach for mixed data clustering in EDM.

6.2. Summary of Research Findings

This research has investigated the use of unsupervised feature selection as the primary solution to the problem of mixed dataset in educational institutions. The investigation revealed that many studies have devoted to predictive approach to educational datasets. The existing methods are mostly tested on labelled data which often poses difficulties because of inaccessibility. Also, they largely depended on numerical feature as the dependent variable, and which is a significant obstacle when the data consist of both numerical and categorical values. In addressing these challenges, this research aimed to pursue the following objectives;

1. To identify existing clustering approaches for treating mixed data in EDM.
2. To propose an alternative approach to unsupervised feature selection for mixed data clustering.
3. To evaluate the proposed approach with existing partition-based clustering methods for mixed datasets.

In achieving the first objective, this thesis conducted a systematic literature review as elaborated in Chapter 2 to identify; (1) the adopted clustering algorithms in EDM, (2) the methods used in measuring the relevance of these clustering algorithms, and (3) the different outcomes, that the clustering (specifically partition based) algorithms provide.

The findings show that mixed data clustering in EDM can be grouped into three categories; namely, numeric data clustering, categorical data clustering, and mixed data clustering. Several

studies are devoted to numeric and categorical data clustering in EDM. See section 2.7.1 for details. Far less research is conducted in mixed data clustering for EDM as discussed in section 2.7.2

Moreover, the findings have shown that a majority of the partition-based clustering works are in the context of e-learning, which is coherent, because, e-learning is operated by software applications that collect interaction and usage data through the e-learning system. This dataset can then be pre-processed and analysed to answer questions. The findings also reported that fewer studies exist where an explicit software application is not used, for instance, classroom decoration or learning styles or the impact of school facilities on student learning performance.

This research then proposes a novel feature selection algorithm for mixed data clustering. To evaluate the significance of the proposed approach, this research utilizes a real-world large school panel level dataset obtained from the District Information System for School Education (DISE), a venture of the Government of India and Ministry of Education, India. This work chose the data from the state of Delhi for the academic year of 2012 to 2013. The dataset comprised of 1,469 primary level schools with 133 mixed data variables.

The second objective of this research is to evaluate the proposed approach on both educational and non-educational datasets of mixed-type. This is because initial studies have realized the over-dependency of the existing approaches solely on numerical educational datasets. This made the approaches work particularly well when numerical data exist, which is a significant hurdle in the construction as a new mixed data clustering system. Since performing these approaches depend on the numerical data distribution, they cease to yield optimum solutions when entrusted with mixed-data types.

Therefore, to address this problem, this research developed a mixed-data feature selection approach that depends upon the information content of a variable, be it numerical or categorical. In the proposed approach, the dataset is separated into numerical and nominal variables. The numerical features are normalized on median and the distance between numerical features n_1 and n_2 is

calculated via Manhattan distance. The nominal features are normalized on frequency centroid. Then, association between features of high variance is computed to ensure the inherent correlations are preserved. Such features are extracted and passed into the modified Gower distance function discussed in section 4.4.

To achieve the third objective, publicly available mixed datasets were used for conducting extensive experiments to test the effectiveness and efficiency of the proposed approach. The experiments have indicated the suggested method has yielded a major improvement over the existing benchmarked algorithms with reference to attainment of significant and consistent groups.

This research is significant as it unveiled an alternative approach to analysing mixed data types of both numeric and nominal. With the proposed method, developers and researchers can build systems to group mixed data without the prerequisite for data conversion from one type to another to suit an algorithm requirement. Also, with the proposed approach researchers need not attain a reasonable level of expertise to pre-process data in future.

The contribution of this research to existing literature are: (1) identification and leveraging the advantages of mixed data clustering in EDM; (2) proposing an alternative approach for grouping different data types that does not require a priori data conversion; (3) maintaining association between variables of different types to yield pure clusters.

6.3. Limitations of the Study

The proposed approach is dependent on the distribution of data and the number of factor levels in a categorical variable. In general, the distribution of numerical data is associated to theσ_i, the average standard aberration of the data points in a group *c*. In practice, $\sigma_i$can be used as a navigation approach to ordain the number of possible groups in the dataset. Although as σ_i is not known before grouping, the overall average standard aberration σ for numerical features of all σ_i can be used. The σ_i can be calculated from the preceding clustering results in an iterative algorithm.

6.4. Recommendation for Future Work

Even though much research has been attracted towards EDM but most of it is directed towards the supervised or classification direction. Moreover, much of this research only considers univariate data types. There is an urgent requirement to pre-process mixed data types especially in an educational setting. The EDM community need to establish a platform to discuss the best-practice guidelines for evaluating mixed data clustering approaches. The discussion should focus on;

1. Providing an open-source educational data sets such as student interaction with school facilities, student and teacher demographic details including parent's occupation and income records. This will facilitate researchers and policy makers to better understand the dynamics between student, teacher, parents and schools.
2. There should be benchmarked educational datasets as well as open sourced algorithms that can be used to calibrate the efficiency of the proposed methods.

Furthermore, while this research has proposed an approach to analyse mixed dataset in educational environment, it is pertinent for the researchers to perform a series of comparisons between the two scenarios to identify the settings conducive for yielding pure clusters in future studies.

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.

Aher, S. B., & Lobo, L. (2012, August). Applicability of data mining algorithms for recommendation system in e-learning. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 1034-1040). ACM.

Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. IEEE Access, 7, 31883-31902.

Ahmad, A., & Dey, L. (2005, December). Algorithm for fuzzy clustering of mixed data with numeric and categorical attributes. In *International Conference on Distributed Computing and Internet Technology* (pp. 561-572). Springer, Berlin, Heidelberg.

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, *63*(2), 503-527.

Ahmad, A., & Dey, L. (2011). A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. *Pattern Recognition Letters, 32*(7), 1062-1069.

Ahmad, A., & Hashmi, S. (2016). K-Harmonic means type clustering algorithm for mixed datasets. *Applied Soft Computing, 48*, 39-49.

A1-Daoud, M.B. & Roberts, S.A (1996). New methods for the initialisation of clusters. Pattern Recognition Letters, 17, 451-455.

Al-Daoud, M.B (2005). A new algorithm for cluster initialization. World Academy of Science, Engineering and Technology, 4, 74-76.

Alahakoon, D., Halgamuge, S. K., & Srinivasan, B. (2000). Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Transactions on Neural Networks, 11*(3), 601-614.

Aloise, D., Deshpande, A., Hansen, P., & Popat, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. Machine learning, 75(2), 245-248.

Almeda, M. V., Scupelli, P., Baker, R. S., Weber, M., & Fisher, A. (2014, March). Clustering of design decisions in classroom visual displays. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (pp. 44-48). ACM.

Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences, 97(18), 10101-10106.

Amershi, S., & Conati, C. (2009). Combining unsupervised and supervised classification to build user models for exploratory. *Journal of Educational Data Mining*, *1*(1), 18-71.

Anaya, A. R., & Boticario, J. G. (2009, April). Clustering learners according to their collaboration. In *2009 13th International Conference on Computer Supported Cooperative Work in Design* (pp. 540-545). IEEE.

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, *60*(3), 383-398.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PéRez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1), 243-256.

Arthur, D. & Vassilvitskii, S.(2007). k-means++ : the advantages of careful seeding. In Proceedings of the 18th Annual ACM-SIAM symposium on Discrete algorithms, 1027-1035.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, 177-194.

Astrahan, M. M. (1970). Speech analysis by clustering, or the hyperphoneme method (No. AIM-124). Department of Computer Science, Stanford Univesity CA

Azam, M., & Saing, C. H. (2017). Assessing the impact of district primary education program in India. *Review of Development Economics, 21*(4), 1113-1131.

Babu, G. P., & Murty, M. N. (1993). A near-optimal initial seed value selection in k-means means algorithm using a genetic algorithm. Pattern recognition letters, 14(10), 763-769.

Baker, R. S. (2010). Data mining for education. *International Encyclopedia of Education, 7*, 112-118.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004, August). Detecting student misuse of intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems* (pp. 531-540). Springer, Berlin, Heidelberg.

Baker, R. S., Gowda, S. M., Corbett, A. T., & Ocumpaugh, J. (2012, June). Towards automatically detecting whether student learning is shallow. In *International Conference on Intelligent Tutoring Systems* (pp. 444-453). Springer, Berlin, Heidelberg.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3-17.

Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. Behavioral science, 12(2), 153-155.

Barcelo-Rico, F., & Diez, J.-L. (2012). Geometrical codification for clustering mixed categorical and numerical databases. *Journal of Intelligent Information Systems, 39*(1), 167-185.

Beck, J. E., & Woolf, B. P. (2000, June). High-level student modeling with machine learning. In *International Conference on Intelligent Tutoring Systems* (pp. 584-593). Springer, Berlin, Heidelberg.

Belkin, M., & Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In Advances in neural information processing systems (pp. 585-591).

Bharti, K. K., Shukla, S., & Jain, S. (2010). Intrusion detection using clustering. *Proceeding of the Association of Counseling Center Training Agencies (ACCTA)*, *1*.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719-725.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*: Springer.

Böhm, C., Goebl, S., Oswald, A., Plant, C., Plavinski, M., & Wackersreuther, B. (2010, June). Integrative parameter-free clustering of data with mixed type attributes. In *Pacific-asia*

conference on knowledge discovery and data mining (pp. 38-47). Springer, Berlin, Heidelberg.

Boriah, S., Chandola, V., & Kumar, V. (2008, April). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining* (pp. 243-254). Society for Industrial and Applied Mathematics.

Bradley, P.S. and Fayyad, U.M. (1998) Refining initial points for k-means clustering. *In the 15th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc, San Francisco*.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104).

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., & Dougherty, E. R. (2007). Model-based evaluation of clustering validation measures. Pattern Recognition, 40(3), 807-824.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing, 300*, 70-79.

Cao, F., Liang, J. and Bai, L. A new initialization method for categorical data clustering. Expert Systems with Applications, 36, 10223-10228 (2009).

Cao, F., Liang, J., & Jiang, G. (2009). An initialization method for the K-Means algorithm using neighborhood model. Computers & Mathematics with Applications, 58(3), 474-483.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods, 3*(1), 1-27.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *Educause Review, 42*(4), 40.

Carpenter, G. A., & Grossberg, S. (2010). *Adaptive resonance theory*: Springer.

Celebi, M.E., Kingravi, H.A. and Vela, P.A.(2013) A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications, 40, 200-210.

Chae, S.-S., Kim, J.-M., & Yang, W.-Y. (2006). Cluster analysis with balancing weight on mixed-type data. *Communications for Statistical Applications and Methods, 13*(3), 719-732.

Chang, W. C., Wang, T. H., & Li, M. F. (2010). Learning Ability Clustering in Collaborative Learning. *Journal of Software, 5*(12), 1363-1370.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2012). NbClust Package: finding the relevant number of clusters in a dataset. J. Stat. Softw.

Chatzis, S. P. (2011). A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. *Expert Systems With Applications, 38*(7), 8684-8689.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., & Freeman, D. (1988). Autoclass: A Bayesian classification system *Machine Learning Proceedings 1988* (pp. 54-64): Elsevier.

Chen, C. M., Chen, Y. Y., & Liu, C. Y. (2007). Learning performance assessment approach using web-based learning portfolios for e-learning systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *37*(6), 1349-1359.

Chen, C. M., Li, C. Y., Chan, T. Y., Jong, B. S., & Lin, T. W. (2007, October). Diagnosis of students' online learning portfolios. In *2007 37th Annual Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports* (pp. T3D-17). IEEE.

Chen, H. M., & Cooper, M. D. (2001). Using clustering techniques to detect usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology, 52*(11), 888-904.

Chen, J.-Y., & He, H.-H. (2016). A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data. *Information Sciences, 345*, 271-293.

Chen, N., & Marques, N. C. (2005, December). An extension of self-organizing maps to categorical data. In *Portuguese Conference on Artificial Intelligence* (pp. 304-313). Springer, Berlin, Heidelberg.

Chen, J., Huang, K., Wang, F., & Wang, H. (2009, October). E-learning behavior analysis based on fuzzy clustering. In *2009 Third International Conference on Genetic and Evolutionary Computing* (pp. 863-866). IEEE.

Cheng, Y.-M., & Leu, S.-S. (2009). Constraint-based clustering and its applications in construction management. *Expert Systems with Applications, 36*(3), 5761-5767.

Chi, C. C., Kuo, C. H., Lu, M. Y., & Tsao, N. L. (2008, July). Concept-based pages recommendation by using cluster algorithm. In *2008 Eighth IEEE International Conference on Advanced Learning Technologies* (pp. 298-300). IEEE.

Chiodi, M. (1990). A partition type method for clustering mixed data. *Rivista di statistica applicata, 2*, 135-147.

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001, August). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data mining* (pp. 263-268). ACM.

Cobo, G., García-Solórzano, D., Morán, J. A., Santamaría, E., Monzo, C., & Melenchón, J. (2012, April). Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 248-251). ACM.

Cobo, G., García-Solórzano, D., Santamaria, E., Morán, J. A., Melenchón, J., & Monzo, C. (2011). Modeling Students' Activity in Online Discussion Forums: A Strategy based on Time Series and Agglomerative Hierarchical Clustering. In *EDM* (pp. 253-258).

Dash, M., Liu, H., & Yao, J. (1997). Dimensionality reduction of unsupervised data. In Proceedings ninth ieee international conference on tools with artificial intelligence (pp. 532-539). IEEE.

David, G., & Averbuch, A. (2012). SpectralCAT: Categorical spectral clustering of numerical and nominal data. *Pattern Recognition, 45*(1), 416-433.

Dean, J., & Ghemawat, S. (2010). MapReduce: a flexible data processing tool. *Communications of the ACM, 53*(1), 72-77.

Del Coso, C., Fustes, D., Dafonte, C., Nóvoa, F. J., Rodríguez-Pedreira, J. M., & Arcay, B. (2015). Mixing numerical and categorical data in a Self-Organizing Map by means of frequency neurons. *Applied Soft Computing, 36*, 246-254.

Dharmarajan, A., & Velmurugan, T. (2013, December). Applications of partition based clustering algorithms: A survey. In *2013 IEEE International Conference on Computational Intelligence and Computing Research* (pp. 1-5). IEEE.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3), 297-302.

Di Ciaccio, A. (2001). MIXISO: A non-hierarchical clustering method for mixed-mode data. In *Advances in Classification and Data Analysis* (pp. 27-34). Springer, Berlin, Heidelberg.

Ding, S., Du, M., Sun, T., Xu, X., & Xue, Y. (2017). An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood. *Knowledge-Based Systems, 133*, 294-313.

Doring, C., Borgelt, C., & Kruse, R. (2004, June). Fuzzy clustering of quantitative and qualitative data. In *IEEE Annual Meeting of the Fuzzy Information, 2004. Processing NAFIPS'04.* (Vol. 1, pp. 84-89). IEEE.

Dradilova, P., Martinovic, J., Slaninová, K., & Snásel, V. (2008). Analysis of Relations in eLearning (Vol. 3, pp. 373–376).

Du, M., Ding, S., & Xue, Y. (2017). A novel density peaks clustering algorithm for mixed data. *Pattern Recognition Letters, 97*, 46-53.

Duan, B., Han, L., Gou, Z., Yang, Y., & Chen, S. (2019). Clustering Mixed Data Based on Density Peaks and Stacked Denoising Autoencoders. *Symmetry, 11*(2), 163.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.

Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering, 5*(2), 112.

Dutt, A., Ismail, M. A. B., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access, 5*, 15991-16005.

El-Sonbaty, Y., & Ismail, M. A. (1998). Fuzzy clustering for symbolic data. *IEEE Transactions on Fuzzy Systems, 6*(2), 195-204.

Elavarasi, S. A., Akilandeswari, J., & Sathiyabhama, B. (2011). A survey on partition clustering algorithms. *International Journal of Enterprise Computing and Business Systems*, *1*(1).

Eranki, K. L., & Moudgalya, K. M. (2012, July). Evaluation of web based behavioral interventions using spoken tutorials. In *2012 IEEE Fourth International Conference on Technology for Education* (pp. 38-45). IEEE.

Everitt, B. S. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters, 6*(5), 305-309.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning, 2*(2), 139-172.

Forgy, E.W. (1965), *Cluster analysis of multivariate data: efficiency vs interpretability of classifications*. Biometrics, 21, 768- 769

Foss, A. H., Markatou, M., & Ray, B. (2019). Distance Metrics and Clustering Methods for Mixed-type Data. *International Statistical Review*, *87*(1), 80-109.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972-976.

Furao, S., & Hasegawa, O. (2006). An incremental network for on-line unsupervised classification and topology learning. *Neural Networks, 19*(1), 90-106.

García, E., Romero, C., Ventura, S., & de Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education, 14*(2), 77-88.

Garcia, S., Luengo, J., Sáez, J. A., Lopez, V., & Herrera, F. (2012). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 25(4), 734-750.

Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial intelligence, 40*(1-3), 11-61.

Ghorbani, F., & Montazer, G. A. (2012, February). Learners grouping improvement in e-learning environment using fuzzy inspired PSO method. In *6th National and 3rd International Conference of E-Learning and E-Teaching* (pp. 65-70). IEEE.

Gluck, M. (1985). Information, uncertainty and the utility of categories. In *Proceedings of the Seventh Annual Conf. on Cognitive Science Society, 1985*.

Goodall, D. W. (1966). A new similarity index based on probability. *Biometrics*, 882-907.

Gonzales, T.F (1985). Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38, 293-306.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.

Grossberg, S. (2013). Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks, 37*, 1-47.

Gu, H., Zhu, H., Cui, Y., Si, F., Xue, R., Xi, H., & Zhang, J. (2018). Optimized scheme in coal-fired boiler combustion based on information entropy and modified K-prototypes algorithm. *Results in Physics, 9*, 1262-1274.

Gu, Q., Li, Z., & Han, J. (2012). Generalized fisher score for feature selection. A*rXiv Preprint arXiv:1202.3725*.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1157-1182.

Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction *Feature Extraction: Foundations and Applications* (pp. 1-25). Berlin, Heidelberg: Springer.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). Clustering algorithms and validity measures. *Proceedings Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM 2001*, 3-22. doi: 10.1109/ssdm.2001.938534

Hall, M., Witten, I., & Frank, E. (2011). Data mining: Practical machine learning tools and techniques. *Kaufmann, Burlington*.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. & Tatham, R. L. 1998. Multivariate data analysis, Upper Saddle River, New York, Pearson.

Han, J., Cai, Y., & Cercone, N. (1993). Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Engineering, 5*(1), 29-40.

Han, J., & Fu, Y. (1994, July). Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases. In *KDD workshop* (pp. 157-168).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 100-108.

He, X., Cai, D., & Niyogi, P. (2006). Laplacian score for feature selection. In Advances in neural information processing systems (pp. 507-514).

He, Z., Xu, X., & Deng, S. (2002). Squeezer: an efficient algorithm for clustering categorical data. *Journal of Computer Science and Technology, 17*(5), 611-624.

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2015). *Handbook of Cluster Analysis*: CRC Press.

Honda, K., & Ichihashi, H. (2005). Regularized linear fuzzy clustering and probabilistic PCA mixture models. *IEEE Transactions on fuzzy systems, 13*(4), 508-516.

Hoppe, H. (2003). A web-based tutoring tool with mining facilities to improve learning and teaching. *Artificial Intelligence In Education: Shaping the Future of Learning Through Intelligent Technologies, 97*(201), 49.

Hsu, C.-C. (2006). Generalizing self-organizing map for categorical data. *IEEE transactions on Neural Networks, 17*(2), 294-304.

Hsu, C.-C., Chen, C.-L., & Su, Y.-W. (2007). Hierarchical clustering of mixed data based on distance hierarchy. *Information Sciences, 177*(20), 4474-4492.

Hsu, C.-C., & Chen, Y.-C. (2007). Mining of mixed data with application to catalog marketing. *Expert Systems with Applications, 32*(1), 12-23.

Hsu, C.-C., & Huang, Y.-P. (2008). Incremental clustering of mixed data based on distance hierarchy. *Expert Systems with Applications, 35*(3), 1177-1185.

Hsu, C., & Lin, S. (2006). Visualized analysis of multivariate mixed-type data via an extended self-organizing map. In *The 6th International Conference on Information Technology and Applications (ICITA 2009)* (pp. 218-223).

Huang, C. T., Lin, W. T., Wang, S. T., & Wang, W. S. (2009). Planning of educational training courses by data mining: Using China Motor Corporation as an example. *Expert Systems with Applications, 36*(3), 7199-7209.

Huang, D., & Chow, T. W. (2005). Effective feature selection scheme using mutual information. *Neurocomputing, 63*, 325-343.

Huang, Z. (1997, February). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st Pacific-Asia Conference On Knowledge Discovery and Data Mining,(PAKDD)* (pp. 21-34).

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery, 2*(3), 283-304.

Huang, C. M., & Harris, R. W. (1993). A comparison of several vector quantization codebook generation approaches. IEEE Transactions on Image Processing, 2(1), 108-112.

Huang, Z., & Ng, M. K. (1999). A fuzzy K-modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems, 7(4), 446-452.

Hunt, L., & Jorgensen, M. (2011). Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(4), 352-361.

Irani, J., Pise, N., & Phatak, M. (2016). Clustering techniques and the similarity measures used in clustering: A survey. *International Journal of Computer Applications, 134*(7), 9-14.

Ivančević, V., Čeliković, M., & Luković, I. (2012, October). The individual stability of student spatial deployment and its implications. In *2012 International Symposium on Computers in Education (SIIE)* (pp. 1-4). IEEE.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs,NJ, USA: Prentice-Hall, Inc.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

Jancey, R.C. (1966). *Multidimensional group analysis*. Australian Journal of Botany, 14, 127-130.

Jang, H.-J., Kim, B., Kim, J., & Jung, S.-Y. (2018). An Efficient Grid-Based K-Prototypes Algorithm for Sustainable Decision-Making on Spatial Objects. *Sustainability, 10*(8), 2614.

Ji, J., Bai, T., Zhou, C., Ma, C., & Wang, Z. (2013). An improved K-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing, 120*, 590-596.

Ji, J., Pang, W., Zheng, Y., Wang, Z., Ma, Z., & Zhang, L. (2015). A novel cluster center initialization method for the K-prototypes algorithms using centrality and distance. *Applied Mathematics & Information Sciences, 9*(6), 2933.

Jia, H., & Cheung, Y.-M. (2017). Subspace clustering of categorical and numerical data with an unknown number of clusters. *IEEE Transactions on Neural Networks and Learning Systems, 29*(8), 3308-3325.

Jinyin, C., Xiang, L., Haibing, Z., & Xintong, B. (2017). A novel cluster center fast determination clustering algorithm. *Applied Soft Computing, 57*, 539-555.

Kacem, M. A. B. H., N'cir, C. E. B., & Essoussi, N. (2015, October). MapReduce-based K-prototypes clustering method for big data. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1-7). IEEE.

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids. *Finding groups in data: an introduction to cluster analysis, 344*, 68-125.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344): John Wiley & Sons.

Katsavounidis, I., Kuo, C. C. J., & Zhang, Z. (1994). A new initialization technique for generalized Lloyd iteration. IEEE Signal processing letters, 1(10), 144-146.

Khan, S. S., & Ahmad, A. (2003). *Computing initial points using density based multiscale data condensation for clustering categorical data.* Paper presented at the 2nd International Conference on Applied Artificial Intelligence, ICAAI.

Khan, S.S., & Ahmad, A. (2004), *Cluster center initialization algorithm for k-means clustering*. Pattern Recognition Letters, 25, 1293-1302.

Khan, S.S. and Kant, S. (2007). Computation of initial modes for k-modes clustering algorithm using evidence accumulation. *In the 20th international joint conference on Artifical intelligence, Morgan Kaufmann Publishers Inc*, San Francisco, 2007.

Khan, S. S., & Ahmad, A. (2013). Cluster center initialization algorithm for K-modes clustering. *Expert Systems with Applications, 40*(18), 7444-7456.

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and Software Technology, 51*(1), 7-15.

Kumar, V, Chhabra, J. & Kumar, D (2011). Initializing cluster center for k-means using biogeography based optimization. In Advances in Computing, Communication and Control. Springer, Berlin, 448-456.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*(1), 59-69.

Lakshmi, K., Shanthi, S., & Parvathavarthini, S. (2018). Clustering mixed datasets using K-prototypes algorithm based on crow-search optimization *Developments and Trends in Intelligent Technologies and Smart Systems* (pp. 191-210): IGI Global.

Lam, D., Wei, M., & Wunsch, D. (2015). Clustering data of mixed categorical and numerical type with unsupervised feature learning. *IEEE Access, 3*, 1605-1613.

Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. The computer journal, 9(4), 373-380.

Lawrence, C., & Krzanowski, W. J. (1996). Mixture separation for mixed-mode data. *Statistics and Computing, 6*(1), 85-92.

Li, C., & Biswas, G. (2002). Unsupervised learning with mixed numeric and nominal data. *IEEE Transactions on Knowledge & Data Engineering*(4), 673-690.

Li, C., & Yoo, J. (2006, March). Modeling student online learning using clustering. In *Proceedings of the 44th annual Southeast Regional Conference* (pp. 186-191). ACM.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature selection: A data perspective. *ACM Computing Surveys (CSUR), 50*(6), 94.

Liang, J., Chin, K.-S., Dang, C., & Yam, R. C. (2002). A new method for measuring uncertainty and fuzziness in rough set theory. *International Journal of General Systems, 31*(4), 331-342.

Liang, J., Zhao, X., Li, D., Cao, F., & Dang, C. (2012). Determining the number of clusters using information entropy for mixed data. *Pattern Recognition, 45*(6), 2251-2265.

Lim, J., Jun, J., Kim, S. H., & McLeod, D. (2012). *A framework for clustering mixed attribute type datasets.* Paper presented at the Proceedings of the 4th International Conference on Emerging Databases.

Lin, D. (1998). An information-theoretic definition of similarity. In *Icml* (Vol. 98, No. 1998, pp. 296-304).

Lin, S.-H. (2012). Data mining for student retention management. *Journal of Computing Sciences in Colleges, 27*(4), 92-99.

Liu, S., & d'Aquin, M. (2017, April). Unsupervised learning for understanding student achievement in a distance learning setting. In *2017 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1373-1377). IEEE.

Liu, S., Zhou, B., Huang, D., & Shen, L. (2017). Clustering mixed data by fast search and find of density peaks. *Mathematical Problems in Engineering, 2017*.

Liu, X., Yang, Q., & He, L. (2017). A novel DBSCAN with entropy and probability for mixed data. *Cluster Computing, 20*(2), 1313-1323.

Jing, L. (2004). Data Mining Applications in Higher Education. *Executive report. SPSS Inc. DMHEWP-1004*.

Ji, J., Pang, W., Zheng, Y., Wang, Z., Ma, Z., & Zhang, L. (2015). *A novel cluster center initialization method for the k-prototypes algorithms using centrality and distance*. Applied Mathematics & Information Sciences, 9(6), 2933.

Linde, Y., Buzo, A., & Gray, R. (1980). An algorithm for vector quantizer design. IEEE Transactions on communications, 28(1), 84-95.

Luo, H., Kong, F., & Li, Y. (2006, August). Clustering mixed data based on evidence accumulation. In *International Conference on Advanced Data Mining and Applications* (pp. 348-355). Springer, Berlin, Heidelberg.

MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. *In the fifth Berkeley Symposium on Mathematical Statistics and Probability*

Manikandan, C., Sundaram, A. M., & Babu, M. M. (2006, December). Collaborative E-learning for remote education; an approach for realizing pervasive learning environments. In *2006 International Conference on Information and Automation* (pp. 274-278). IEEE.

Manyika, J. (2011). Big data: The next frontier for innovation, competition, and productivity. *http://www. Mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation*.

Marbac, M., Biernacki, C., & Vandewalle, V. (2017). Model-based clustering of Gaussian copulas for mixed data. *Communications in Statistics-Theory and Methods, 46*(23), 11635-11656.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems, 33*(1), 107-124.

McKusick, K., & Thompson, K. (1990). Cobweb/3: A portable implementation.

McLachlan, G., & Krishnan, T. (2007). *The EM algorithm and extensions* (Vol. 382): John Wiley & Sons.

McParland, D., & Gormley, I. C. (2016). Model based clustering for mixed data: clustMD. *Advances in Data Analysis and Classification, 10*(2), 155-169.

Melnykov, V., & Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys, 4*, 80-116.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45(3), 325-342.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159-179.

Mirkin, B. (2001). Reinterpreting the category utility function. *Machine Learning, 45*(2), 219-228.

Mitra, P., Murthy, C.A., Pal, S.K. (2002). *Density-based multiscale data condensation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 734-747.

Modha, D. S., & Spangler, W. S. (2003). Feature weighting in k-means clustering. *Machine Learning, 52*(3), 217-237.

Mohsin, M. F. M., Norwawi, N. M., Hibadullah, C. F., & Wahab, M. H. A. (2010). *Mining the student programming performance using rough set*. Paper presented at the Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation. The Computer Journal, 20(4), 359-363.

Morlini, I., & Zani, S. (2010). Comparing approaches for clustering mixed mode data: an application in marketing research *Data Analysis and Classification* (pp. 49-57): Springer.

Moustaki, I., & Papageorgiou, I. (2005). Latent class models for mixed variables with applications in Archaeometry. *Computational Statistics & Data Analysis, 48*(3), 659-675.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (pp. 849-856).

Niijima, S., & Okuno, Y. (2008). Laplacian linear discriminant analysis approach to unsupervised feature selection. IEEE/ACM transactions on computational biology and bioinformatics, 6(4), 605-614.

Niu, K., Niu, Z., Su, Y., Wang, C., Lu, H., & Guan, J. (2015). A coupled user clustering algorithm based on mixed data for web-based learning systems. *Mathematical Problems in Engineering, 2015*.

Noorbehbahani, F., Mousavi, S. R., & Mirzaei, A. (2015). An incremental mixed data clustering method using a new distance measure. *Soft Computing, 19*(3), 731-743.

Norouzi, M., Fleet, D. J., & Salakhutdinov, R. R. (2012). Hamming distance metric learning. In *Advances in Neural Information Processing Systems* (pp. 1061-1069).

Onoda, T., Sakai, M., & Yamada, S. (2012). Careful seeding method based on independent components analysis for k-means clustering. Journal of Emerging Technologies in Web Intelligence, 4(1), 51-59.

Parack, S., Zahid, Z., & Merchant, F. (2012, January). Application of data mining in educational databases for predicting academic trends and patterns. In *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)* (pp. 1-4). IEEE.

Parack, S., Zahid, Z., & Merchant, F. (2012b). Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns. *2012 IEEE International Conference on Technology Enhanced Education (Ictee 2012)*, 1-4.

Pathak, A., & Pal, N. R. (2016). Clustering of mixed data by integrating fuzzy, probabilistic, and collaborative clustering framework. *International Journal of Fuzzy Systems, 18*(3), 339-348.

Pedersen, T., & Kulkarni, A. (2006, June). Automatic cluster stopping with criterion functions and the gap statistic. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations (pp. 276-279).

Pedrycz, W. (2002). Collaborative fuzzy clustering. *Pattern Recognition Letters, 23*(14), 1675-1686.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaiane, O. R. (2009). Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE Transactions on Knowledge and Data Engineering, 21*(6), 759-772.

Philip, G., & Ottaway, B. (1983). Mixed data cluster analysis: an illustration using Cypriot hooked-tang weapons. *Archaeometry, 25*(2), 119-133.

Plant, C., & Böhm, C. (2011, August). Inconco: interpretable clustering of numerical and categorical objects. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1127-1135). ACM.

Pfitzner, D., Leibbrandt, R., & Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. Knowledge and Information Systems, 19(3), 361.

Pizzuti, C., Talia, D., & Vonella, G. (1999). A divisive initialisation method for clustering algorithms. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 484-491). Springer, Berlin, Heidelberg.

Podani, J. (1999). Extending Gower's general coefficient of similarity to ordinal characters. *Taxon, 48*(2), 331-340.

Prasad, D. H., & Punithavalli, M. (2012, March). An Integrated Framework for Mixed Data Clustering Using Growing Hierarchical Self-Organizing Map (GHSOM). In *International Conference on Mathematical Modelling and Scientific Computation* (pp. 471-479). Springer, Berlin, Heidelberg.

Rahman, M. A., & Islam, M. Z. (2012, December). CRUDAW: a novel fuzzy technique for clustering records following user defined attribute weights. In *Proceedings of the Tenth Australasian Data Mining Conference-Volume 134* (pp. 27-41). Australian Computer Society, Inc.

Rajan, V., & Bhattacharya, S. (2016, July). Dependency Clustering of Mixed Data with Gaussian Mixture Copulas. In *IJCAI* (pp. 1967-1973).

Ranjan, J., & Malik, K. (2007). Effective educational process: a data-mining approach. *Vine*, *37*(4), 502-515.

Rashid, N. A., Taib, M. N., Lias, S., Sulaiman, N., Murat, Z. H., & Kadir, R. S. S. A. (2011). Learners' Learning Style Classification related to IQ and Stress based on EEG. *Procedia-Social and Behavioral Sciences*, *29*, 1061-1070.

Redmond, S.J. & Heneghan, C (2007). A method for initialising the k-means clustering algorithm using kd-trees. Pattern Recognition Letters, 28, 965-973.

Reich, Y., & Fenves, S. J. (1991). The formation and use of abstract concepts in design *Concept Formation* (pp. 323-353): Elsevier.

Ren, M., Liu, P., Wang, Z., & Pan, X. (2016, August). An improved mixed-type data based kernel clustering algorithm. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 1205-1209). IEEE.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Rhodes, F. H. T. (2001). *The creation of the future: The role of the American university*: Cornell University Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465-471.

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science, 344*(6191), 1492-1496.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. PloS one, 14(1).

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68*, 458-472.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Roy, D. K., & Sharma, L. K. (2010). Genetic k-Means clustering algorithm for mixed numeric and categorical data sets. *International Journal of Artificial Intelligence & Applications, 1*(2), 23-28.

Saâdaoui, F., Bertrand, P. R., Boudet, G., Rouffiac, K., Dutheil, F., & Chamoux, A. (2015). A dimensionally reduced clustering methodology for heterogeneous occupational medicine data mining. *IEEE Transactions on Nanobioscience, 14*(7), 707-715.

Solorio-Fernández, S., Martínez-Trinidad, J. F., & Carrasco-Ochoa, J. A. (2017). A new unsupervised spectral feature selection method for mixed data: a filter approach. *Pattern Recognition, 72*, 314-326.

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42-47). IEEE.

Salazar, A., Gosalbez, J., Bosch, I., Miralles, R., & Vergara, L. (2004, June). A case study of knowledge discovery on academic achievement, student desertion and student retention. In *ITRE 2004. 2nd International Conference Information Technology: Research and Education* (pp. 150-154). IEEE.

Salvador, S., & Chan, P. (2005). Learning states and rules for detecting anomalies in time series. Applied Intelligence, 23(3), 241-255.

Sangam, R. S., & Om, H. (2018). An equi-biased K-prototypes algorithm for clustering mixed-type data. *Sādhanā, 43*(3), 37.

Santos, O. C., & Boticario, J. G. (2010). Modeling recommendations for the educational domain. *Procedia Computer Science, 1*(2), 2793-2800.

Sardareh, S. A., Aghabozorgi, S., & Dutt, A. (2014). *Reflective Dialogues and Students' Problem Solving Ability Analysis Using Clustering*. Paper presented at the The 3rd International Conference on Computer Engineering and Mathematical Sciences (ICCEMS 2014), Langkawi, Malaysia.

Schwarz, B. B., & Glassner, A. (2003). The blind and the paralytic: Supporting argumentation in everyday and scientific issues *Arguing to learn* (pp. 227-260): Springer.

Sharpe, D. (2015). Your chi-square test is statistically significant: Now what? *Practical Assessment, Research & Evaluation, 20*.

Shekhawat, M., & Sharma, I. (2017, April). A new dissimilarity metric based on density and connectivity. In *2017 International Conference on Communication and Signal Processing (ICCSP)* (pp. 0923-0927). IEEE.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008, August). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 614-622). ACM.

Shih, M.-Y., Jheng, J.-W., & Lai, L.-F. (2010). A two-step method for clustering mixed categroical and numeric data. *Tamkang Journal of Science and Engineering, 13*(1), 11-19.

Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology, 9*(4).

Sowjanya, A., & Shashi, M. (2011). A cluster feature-based incremental clustering approach to mixed data. *Journal of Computer Science, 7*(12), 1875.

Sparck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1), 11-21. Later: Journal of Documentation, 60(5) (2002), 493-502.

Späth, H. (1977). Computational experiences with the exchange method: Applied to four commonly used partitioning cluster analysis criteria. European Journal of Operational Research, 1(1), 23-31.

Su, T., & Dy, J. G. (2007). In search of deterministic methods for initializing K-means and Gaussian mixture clustering. Intelligent Data Analysis, 11(4), 319-338.

Sun, Y, Zhu, Q.M. & Chen, Z.X (2002). An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recognition Letters, 23, 875-884.

Šulc, Z. "Similarity Measures for nominal variable clustering." The 8th International Days of Statistics and Economics. Slaný: Melandrium (2014): 1536-1545

Šulc, Z., Matejka, M., & Procházka, J. (2016). Modifications of the Gower similarity coefficient. In *The 19th Conference of Applications of Mathematics and Statistics in Economics, Banská Bystrica. Available at: http://amse. umb. sk/proceedings/SulcProchazkaMatejka.pdf*.

Szepannek, G. (2018). clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *The R Journal, 10*(2), 200-208.

Tai, W.-S., & Hsu, C.-C. (2012). Growing Self-Organizing Map with cross insert for mixed-type data clustering. *Applied Soft Computing, 12*(9), 2856-2866.

Tair, M. M. T., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: A case study. *International Journal of Information and Communication Technology Research, 2*(2).

Tan, P. N., Steinbach, M., & Kumar, V. (2013). Data mining cluster analysis: Basic concepts and algorithms.

Tang, T. Y., & McCalla, G. (2005). Smart recommendation for an evolving e-learning system: Architecture and experiment. *International Journal on ELearning, 4*(1), 105.

Tekumalla, L. S., Rajan, V., & Bhattacharyya, C. (2017). Vine copulas for mixed data: multi-view clustering for mixed data beyond meta-Gaussian dependencies. *Machine Learning, 106*(9-10), 1331-1357.

Tian, F., Wang, S., Zheng, C., & Zheng, Q. (2008, April). Research on e-learner personality grouping based on fuzzy clustering analysis. In *2008 12th International Conference on Computer Supported Cooperative Work in Design* (pp. 1035-1040). IEEE.

Tou, J. T., & Gonzalez, R. C. (1974). Pattern recognition principles.

Trandafili, E., Allkoçi, A., Kajo, E., & Xhuvani, A. (2012, September). Discovery and evaluation of student's profiles with machine learning. In *Proceedings of the Fifth Balkan Conference in Informatics* (pp. 174-179). ACM.

Varshavsky, R., Gottlieb, A., Linial, M., & Horn, D. (2006). Novel unsupervised feature filtering of biological data. Bioinformatics, 22(14), e507-e513.

Valsamidis, S., Kontogiannis, S., Kazanidis, I., Theodosiou, T., & Karakos, A. (2012). A Clustering Methodology of Web Log Data for Learning Management Systems. *Educational Technology & Society, 15*(2), 154-167.

van de Velden, M., Iodice D'Enza, A., & Markos, A. (2019). Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *11*(3), e1456.

Wang, C., Chi, C. H., Zhou, W., & Wong, R. (2015, February). Coupled interdependent attribute analysis on mixed data. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Wang, L., Lei, Y., Zeng, Y., Tong, L., & Yan, B. (2013). Principal feature analysis: A multivariate feature selection method for fMRI data. *Computational and Mathematical Methods in Medicine, 2013*.

Wangchamhan, T., Chiewchanwattana, S., & Sunat, K. (2017). Efficient algorithms based on the k-means and chaotic league championship algorithm for numeric, categorical, and mixed-type data clustering. *Expert Systems with Applications, 90*, 146-167.

Wei, M., Chow, T., & Chan, R. (2015). Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation. *Entropy, 17*(3), 1535-1548.

Weiss, N. A. (2015). *Introductory Statistics* (10th Ed.): Pearson.

Welcome to the Trac Open Source Project. (2014). Retrieved 10 June, 2014, from http://trac.edgewall.org

Wook, M., Yahaya, Y. H., Wahab, N., Isa, M. R. M., Awang, N. F., & Seong, H. Y. (2009, December). Predicting NDUM student's academic performance using data mining techniques. In *2009 Second International Conference on Computer and Electrical Engineering* (Vol. 2, pp. 357-361). IEEE.

Wu, S., Jiang, Q., and Huang, J.Z.(2007). A new initialization method for clustering categorical data. *In the 11th PacificAsia Conference on Knowledge Discovery and Data Mining (PAKDD)*

Yang, M.-S. (1993). A survey of fuzzy clustering. *Mathematical and Computer Modelling, 18*(11), 1-16.

Yi, B., Qiao, H., Yang, F. & Xu, C (2010). An improved initialization center algorithm for k-means clustering. In 2010 International Conference on Computational Intelligence and Software Engineering (CiSE),1-4.

Yin, H. (2002). ViSOM-a novel method for multivariate data projection and structure visualization. *IEEE Transactions on Neural Networks, 13*(1), 237-243.

Zaíane, O. R. (2002, December). Building a recommender agent for e-learning systems. In *International Conference on Computers in Education, 2002. Proceedings.* (pp. 55-59). IEEE.

Zaiane, O. (2001). Web usage mining for a better web-based learning environment.

Zhang, B. (2001, April). Generalized k-harmonic means–dynamic weighting of data in unsupervised learning. In *Proceedings of the 2001 SIAM International Conference on Data Mining* (pp. 1-13). Society for Industrial and Applied Mathematics.

Zhang, D., Ji, M., Yang, J., Zhang, Y., & Xie, F. (2014). A novel cluster validity index for fuzzy clustering based on bipartite modularity. Fuzzy Sets and Systems, 253, 122-137.

Zhang, K., & Gu, X. (2014). An affinity propagation clustering algorithm for mixed numeric and categorical datasets. *Mathematical Problems in Engineering, 2014*.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Record, 25*(2), 103-114.

Zhao, W. D., Dai, W. H., & Tang, C. B. (2007, May). K-centers algorithm for clustering mixed type data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 1140-1147). Springer, Berlin, Heidelberg.

Zhao, X., Liang, J., & Dang, C. (2017). Clustering ensemble selection for categorical data based on internal validity indices. *Pattern Recognition, 69*, 150-168.

Zhao, Z., & Liu, H. (2007, June). Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th international conference on Machine learning (pp. 1151-1157).

Zheng, Q., Ding, J., Du, J., & Tian, F. (2007, April). Assessing method for e-learner clustering. In *2007 11th International Conference on Computer Supported Cooperative Work in Design* (pp. 979-983). IEEE.

Zheng, X., & Jia, Y. (2011, December). A study on educational data clustering approach based on improved particle swarm optimizer. In *2011 IEEE International Symposium on IT in Medicine and Education* (Vol. 2, pp. 442-445). IEEE.

Zheng, Z., Gong, M., Ma, J., Jiao, L., & Wu, Q. (2010, July). Unsupervised evolutionary clustering algorithm for mixed type data. In *IEEE Congress on Evolutionary Computation* (pp. 1-8). IEEE.

Qu, Z., & Wang, X. (2008, December). Application of RS and clustering algorithm in distance education. In *2008 International Workshop on Education Technology and Training & 2008 International Workshop on Geoscience and Remote Sensing* (Vol. 1, pp. 7-10). IEEE.

Zhu, X., Zhang, P., Lin, X., & Shi, Y. (2007, October). Active learning from data streams. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (pp. 757-762). IEEE.

Zorrilla, M. E., Menasalvas, E., Marin, D., Mora, E., & Segovia, J. (2005, February). Web usage mining project for improving web-based learning sites. In *International Conference on Computer Aided Systems Theory* (pp. 205-210). Springer, Berlin, Heidelberg.