A PARTITION-BASED FEATURE SELECTION METHOD FOR MIXED DATA: A FILTER APPROACH

Ashish Dutt[1] and Maizatul Akmar Ismail[2*]

[1,2]Department of Information Systems
Faculty of Computer Science and Information Technology, University of Malaya
50603, Kuala Lumpur, Malaysia

Email: ashish_dutt@siswa.um.edu.my[1], maizatul@um.edu.my[2*](corresponding author)

DOI: https://doi.org/10.22452/mjcs.vol33no2.5

ABSTRACT

Feature selection is fundamentally an optimization problem for selecting relevant features from several alternatives in clustering problems. Though several algorithms have been suggested, however till this day, there has not been any one of those that has been dubbed as the best for every problem scenario. Therefore, researchers continue to strive in developing superior algorithms. Even though clustering process is considered a pre-processing task but what it really does is just dividing the data into groups. In this paper we have attempted an improved distance function to cluster mixed data. A similarity measure for mixed data is Gower distance is adopted and modified to define the similarity between object pairs. A partitional algorithm for mixed data is employed to group similar objects in clusters. The performance of the proposed method has been evaluated on similar mixed and real educational dataset in terms of the silhouette coefficient. Results reveal the effectiveness of this algorithm in unsupervised discovery problems. The proposed algorithm performed better than other clustering algorithms for various datasets.

Keywords: ***Clustering, educational data mining, mixed data, unsupervised feature selection.***

1.0 INTRODUCTION

The rapid development of information systems has witnessed a surge in data generation. Such data often has several features either univariate or multivariate characteristics. To process such datasets is often a challenging task in terms of extracting or selecting prominent features [1]. Curtailing features in a dataset is known as Feature Selection (FS). It is a popular low variance feature reduction method and a data pre-processing method. The method works by selecting an optimum set of features from the original set dependent on a criterion function. It plays a pivotal role in data compression whereby features with low variance are removed. Selecting an appropriate FS method helps in enhancing the algorithm accuracy and eases the interpretation of the results [2].

The FS methods are categorized into several types;

a) FS method that is supervised is applied to training dataset that consist of labelled features. Whereas unsupervised FS method is applied to training dataset that have unlabeled features. The semi-supervised FS method is applied when it has partial labelled features.
b) Depending on the relationship between the features and the learning techniques, the FS method is categorised either into filter, wrapper or embedded model.
c) Depending on the relationship between the learning technique and evaluation criterion, the FS method either be correlation-based, distance based or information-loss based.
d) Depending on the search feature within the feature space, the FS method can be dependent on forward increase, backward deletion, random, and hybrid models.

The research field of data mining in education is concerned with the use of statistical or mathematical models to data generated within the educational environment itself. This environment can either be of a school, college or a university. Information systems in this environment can be record systems for either attendance or examination. The data generated or stored by such a system is often a mix of numerical and categorical features. For instance, a typical student record file 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. In general, the educational datasets are univariate which means its features are either numerical or categorical. In the literature, there are many studies related to supervise feature selection for analysing univariate data type [3-5] in Educational Data Mining

(EDM). To the best of our knowledge, there is not enough discussions on Unsupervised Feature Selection (UFS) for mixed data in EDM. Although it is a known fact that clustering is a preprocessing algorithm. However, it is important to highlight any given clustering algorithm will only divide the data into groups because there is no feature selection involved. Given this approach, we argue that the resulting clusters can often be impure.

We now present the organization of this paper. A review of recent research works on FS and its use in unsupervised learning in EDM context is discussed in section 2.0. In section 3.0, we discuss Gower's distance for measuring the similarity in mixed data and in sub-section 3.3, we have discussed our proposed approach. Thereafter, the proposed approach is applied to a case study of real-life educational dataset, as discussed in section 4.0. Besides the case study, in section 5, we elaborate and test the proposed approach on 5 mixed datasets obtained from University of California Machine Learning repository (UCI ML) as well as on two similar partition-based clustering algorithms. In section 6, we present the discussion and in section 7, we conclude the work.

2.0 LITERATURE REVIEW

We begin this section with an overview of clustering algorithms, where we discuss the types of clustering algorithms. It is followed by an overview of feature selection as a preprocessing step in clustering, where we elaborated on the supervised and unsupervised feature selections. Thereafter, we presented a discussion on EDM with specific reference to clustering and feature selection. The range of published research papers that was considered in this study were between 2016 and 2019. The search string used was, *[("unsupervised feature selection") & ("clustering") & ("educational data mining")], [("supervised feature selection") & ("clustering") & ("educational data mining")] throughout* three databases namely ACM Digital, IEEExplore and Google Scholar.

2.1 An overview of clustering algorithms

Clustering is a process in which patterns are detected within a learning space. A cluster is a conglomerate of objects which have similarities to one another, such as having the same colour or shape. Moreover, clustering is an important unsupervised learning method [6]. While a preprocessing algorithm can significantly reduce the size of a dataset, it is imperative not to lose information when this happens. Hitherto, the literature on classification of clustering methods has not arrived to a consensus. Traditionally, clustering is broadly classified as *hierarchical* and *partitional*. We now provide a brief discussion on them.

2.1.1 Hierarchical based clustering

In hierarchical clustering, every observation holds a membership of a separate group with a well-defined centroid [7]. This centroid contains the median of all the data values within a group, where the objective is to repeatedly merge the nearest clusters with each other progressively as the algorithm hierarchically processes the dataset to yield a nested sequence of partitions. This "bottom-up" approach is also known as *agglomerative clustering.* While hierarchical clustering is a bottom-up approach for data grouping, divisive clustering is rather a "top-down" approach. It begins with an initial huge group of data values, which is broken down into several smaller sub-groups with recursive splits.

2.1.2 Partition based clustering

In contrast to the hierarchical clustering, the partitional clustering algorithm groups the objects into a set of non-overlapping subsets (clusters), such that each data point is precisely within one cluster [7]. The drawback is the initial number of clusters are required to be specified. Given a dataset of N values, the algorithm builds $P(N \geq P)$ data groups, where every group represents a cluster. It groups the data into P clusters where (a) each cluster encapsulates a minimum of one object, and (b) each object is a member of a unique cluster. On the contrary, soft clustering is fuzzy like methods in which the object can yield membership of more than one cluster. Continuing further, the partitioning algorithms advocate to minimize an objective function. For example, k-means and k-medoids methods try to minimize the function $\sum_{i=1}^{p}\sum_{j=1}^{|B_i|} Dist(\mathrm{x}_j, \mathrm{center(i)}), |\mathrm{B}_i|$ is the total count of objects in cluster *i.* One of the most popular partition base algorithms is the k-means that manipulates the average of the data points as its centre within the cluster. Yet another popular partitional clustering method called the k-medoids generates one-level partitioning and non-overlapping spherical shaped clusters.

2.1.2.1 Partition based clustering algorithm for mixed data

In year 1997, researcher Huang [8] proposed the *K-prototypes* algorithm for mixed data clustering which has three phases; the *initial prototypes selection*, *cluster allocation,* and finally the *re-allocation*. In the initial step, a randomized selection of *n* data points as cluster centres is made. Next, squared Euclidean distance metric is applied to compute similarity between attributes of numerical types. Thereafter, distance measurement for categorical attributes is based on their mode. And finally, in the third step which is the reallocation phase, the initial cluster 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 attributes. Notably, for numerical values the distance measured is squared Euclidean distance which is susceptible to high values whereas categorical attributes the distance measured is frequency. Only high frequency valued categorical attributes are considered by this algorithm. The categorical attributes with a lower frequency are discarded that directly leads to information loss. Researchers Ji, Bai, Zhou, Ma, & Wang in 2013, suggested an improved version of the K-prototypes algorithm. Their idea was to determine *distribution centroid* of the categorical features. A distribution centroid is the occurrence frequency of a given feature. There were three problems with this approach. First, they used the mean distance for numeric features which is the same as Huang's [8] approach for distance measured between the numeric features. It is also unclear from that paper the data type of the categorical feature such as ordinal or nominal. Secondly, their approach does not consider the association between features. And finally, the time complexity of their proposed approach is higher $O(k(m+p+Nm-Np)nl$ than the approach that Huang reported. A map reduce based k-prototype approach for big mixed data was proposed by Kacem, N'cir and Essoussi [9]. The drawback with the proposed approach is that it is the same as the Huang's approach, and the only difference being the inclusion of map-reduce algorithm in it to handle big data. In a recent paper by Kumar, Rani and Rao [10] which had suggested a feature scaling approach for handling mixed data using the k-prototypes algorithm. Their idea was to normalise the range of the numerical data. We disputed that approach as range normalization can lead to information loss. Moreover, it is unclear the distance measurement approach for both numerical and categorical data. Also, the researchers have not provided the treatment for categorical data. Another problem with their approach is that it is ambiguous on how the optimum number of partitions was obtained.

2.2 An overview of feature selection in clustering

Feature Selection (FS) is a process to automatically or manually 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 criterion. 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 [11]. 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. In this paper, our focus is on FS method, particularly unsupervised feature selection. The FS methods are categorized into various type;

2.2.1 Supervised Feature Selection (SFS)

SFS is specified for classification type problems. It works by detecting feature correlation with the class label. A SFS method when applied to a dataset, works as $D=(X,C)$, consisting of features $X=\{x_1,x_2,...,x_n\}$ 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 [12], [13].

2.2.2 Unsupervised Feature Selection (UFS)

The UFS methods focus on the natural grouping of data values to enhance the accuracy of clustering by determining an optimum subset of features. The UFS algorithms are simply filter and wrapper based [12], [13].

2.2.2.1 Unsupervised filter model

The filter model is based on the relationship between feature and the class label. It has a lesser computational cost as compared to the wrapper model. In this model, the clustering evaluation criterion has a major role to play because

the clustering algorithm is not utilised for the FS process. This model studies the statistical data properties of the variables for decision making [12], [13].

2.2.2.2 Unsupervised wrapper model

The wrapper model proactively uses a clustering algorithm for evaluating the cluster validity. In this model, the feature subset with maximum performance is adjudged as the final feature subset. Although, the clustering performance of this model is better than the filter model however it is disadvantaged when it comes to computation complexity which is higher than the filter model [12], [13].

2.3 An overview of distance methods and data transformation approaches for mixed data clustering

In this section, we discuss the distance measured used for measuring similarity between numerical and categorical features. A categorical feature is considered nominal when it does not have any order to it. Example of a categorical feature is *colour* which can have entries such as *red, blue, and green, yellow*. There is no order to it, hence it is a nominal feature.

2.3.1 Distance methods for mixed data clustering

In 1971, Gower proposed a mathematical formulae of distance measurements for mixed data [14]. This finding is important to mention as it has laid the 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 $x = 1, 2, ..., n$ (the number of features is denoted by n) and $y = \{1, 2, ..., f\}$ (f is the number of features). Then, the dissimilarity between the objects $a_x = [a_{x1}, a_{x2}, ..., a_{xf}]$ and $a_y = [a_{y1}, a_{y2}, ..., a_{yf}]$ 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 normalized. The ordinal features are rank-ordered and subtracted by 1 and finally range-normalized like the numeric features. This is a simplistic approach. 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 [15]. To overcome this problem, Podani suggested to orderly rank all ordinal features at the initial stage. Features with similar rank were within a close proximity to each other thus not affecting 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 [16] proposed a weight based approach to remedy the problem associated with Gower's approach. Their idea was by assigning weights, it will prevent feature dominance. Pearson's correlation coefficient was used to measure feature similarity for numeric features while Product moment correlation was used to determine the categorical features, which were binary encoded. However, it is not clear the data type of categorical feature, if it was nominal or ordinal. Moreover, the authors had applied the principal component analysis to reduce data dimensionality for obtaining significant features.

Continuing further, in particular we discuss a recent paper by Sulc [17] who presented three modifications for treating nominal data. In the first modification, the authors introduced *"variable entropy"*, in which the concept of weights is used. A higher weight was given to a nominal feature that has a higher variability. The authors asserted that such variables were rare as compared to nominal features with lower variability. In a way, we think that this assertion is incorrect because nominal features with higher variability are actually not *rare* but rather prominent, thus assigned with higher weights. Our inference complies with a similar justification given by Gower [14] on assignment of weights to nominal features. In the second modification, the authors [17] applied the *"Inverse Occurrence Frequency"* concept and assigned greater weights to infrequent mismatches between the nominal features. In the third modification, the authors assigned greater weights to mismatches between nominal features having less number of categories. This approach is similar to the method proposed by Lin in 1988 [18]. This assigned weight takes a value between 0 and 1. It is important to note, that like Gower, the authors proposed modifications will work for two categories in a nominal feature. Also, the authors have not discussed the numeric feature treatment in their proposed modifications unlike the Gower's method where numeric feature treatment was

given. Furthermore, the authors tested their proposed modifications for hierarchical clustering method, namely the two-step cluster analysis and the latent class analysis where Rand Index was used as a cluster evaluation metric.

2.3.2 Data transformation for mixed data clustering

In this section, we present a discussion on existing approaches on mixed data transformation for clustering. First, we will discuss the discretization from numeric feature to categorical and the use of appropriate categorical clustering method. Next, we will discuss numerical coding of categorical features for clustering.

a. Discretization

Discretization of a numerical feature is a widely used method in statistics and machine learning. In this approach, all numeric features are discretized and an applicable clustering method for categorical data is used (e.g. the k-modes algorithm [19]). A possibility of data loss is imminent in the discretization process if inappropriate cut-off points are used [20].

b. 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 [20]. 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 [20].

2.4 Educational Data Mining (EDM)

The information system in a school or college is comprised of either an examination record database or an attendance system. The data stored in such a system is processed and mined to reveal interesting patterns about its users. These patterns can then be analysed to study its users and their interaction with the system. Mining educational data to improve the quality of learning and teaching is called EDM [6].

2.4.1 Clustering algorithms applied in EDM

There exist several studies that have applied supervised feature selection approach to univariate educational data like student examination marks or enrollment count [3-5]. A univariate data type consists of either numerical or categorical data. On the contrary, a mixed dataset consists of both numerical and categorical data types. We found only one study to have addressed this issue. The researchers [21] applied the K-prototypes algorithm to a mixed dataset consisting of student demographics and achievements in a distance learning program. The limitation of this study is that it fell short of examining the feature selection. We believe that this may have been the first attempt to explore the UFS for mixed data in EDM. Further research is warranted 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?"* We postulate that clustering is a preprocessing method that helps in detecting groups by studying the object properties which can be further investigated.

2.4.2 Supervised Feature Selection (SFS) algorithms in EDM

The goal of SFS, is to classify accurately. Table 1 is a review of recent research work on SFS in EDM which depicts that none of the analytical studies reported the data type. Both filter and wrapper based SFS algorithms were used in all studies except for one that utilised human judgement for SFS. All studies reported SFS was helpful in improving the classifier accuracy.

Table 1: Related works on SFS algorithms applied in EDM

S. No.	Reference	SFS algorithms	FS type	Classification method	Data type (univariate/ multivariate)	Evaluation metric	Findings
1.	[22]	Forward semi-supervised	Wrapper	Naïve Bayes	Not given	10-fold cross validation & Cohen's Kappa	Predicting task completion in an Intelligent Tutoring System (ITS)
2.	[23]	ReliefF	Filter	Association rule mining	Not given	Confidence, Lift, Leverage, Conviction	Predicting student success in exams
3.	[24]	Chi-Square base Feature Selection, Information Gain Attribute Evaluation, Principal Components, Relief Attribute	Filter	Decision Tree, Naïve Bayes, j48, Random Forest, Regression Tree, SMO, OneR	Not given	Precision, Recall, F-measure	FS was helpful in improving the predictive accuracy of student performance.
4.	[25]	Naïve Bayes, OneR	Wrapper	Decision Tree, Naïve Bayes, j48, Random Forest, Regression Tree, SMO, OneR	Not given	Precision, Recall, F-measure	FS was helpful in improving the predictive accuracy of student performance.
5.	[26]	Naive Bayes, Decision Tree	None. Human advice sought to determine relevant features	Decision Tree, Naïve Bayes, j48	Not given	Accuracy	Strong correlation found between student exam performance and English and Arabic subjects.
6.	[27]	Linear Discriminate Analysis, Classification and Regression Tree, k-Nearest Neighbor, Support Vector Machine and Random Forest	Learning Vector Quantization	Linear Discriminate Analysis, Classification and Regression Tree, k-Nearest Neighbor, Support Vector Machine and Random Forest	Multivariate	kappa statistics and confusion matrix	FS was helpful in improving the predictive accuracy of student performance

2.4.3 Unsupervised Feature Selection (UFS) algorithms in EDM

The goal of UFS, is to ensure the purity of the cluster. Table 2 presents a review of recent research work on UFS techniques in EDM where none of the analytical studies discussed the data type or the cluster evaluation metric except for one study which had reported the use of a filter based UFS method. There was concurrence in reporting that UFS was helpful in clustering in all the studies. Only one study reported the use of a filter based UFS method but the FS algorithm in this study is actually a feature extraction algorithm, the PCA, which is different from FS.

Table 2: Related works on UFS algorithms applied in EDM

S. No.	Reference	UFS algorithm	FS type	Clustering method	Data type (univariate/ multivariate)	Evaluation metric	Findings
1	[28]	Principal Component Analysis (PCA)-feature extraction method	Filter	K-means. Expectation-Maximization, Hierarchical clustering, DBSCAN	Not given	Not given	Classifying student answers using clustering
2	[29]	Not given	Not given	Average link hierarchical	Not given	Not given	UFS helped identifying affective reactions in educational games.

In reviewing the recent work shown in Table 1 and 2, we find that none of the studies actually discussed the types of data used in their analysis. This is a major problem and cannot be overlooked because in the absence of feature data types, it is difficult to speculate whether or not FS was exercised on numerical or categorical data. Knowing the feature data type is imperative as it ascertains the machine's learning path. Therefore, based on the algorithm type and empirical results shown by studies in Table 1 and 2, we conclude the attribute data type in these studies were numerical in nature. Moreover, filter-based FS methods are more common as compared to wrapper-based FS method in both supervised and unsupervised FS techniques. Apart from that, the evaluation metrics plays a pivotal role in an analytical study. In conclusion, because of the absence of research comprising mixed data clustering in EDM, the desire to bridge this gap was pursued. The purpose of this research is to propose a partitional filter-based FS approach for the mixed data in EDM. In the following sections, we present and discuss the feasibility of our proposed approach.

In this paper, we determine feature relevancy by retaining uncorrelated features and then group them on their structural property. Furthermore, we have attempted to improvise Huang's algorithm by introducing unsupervised feature selection as a precursory step. Our proposed method has two contributions, primarily; it overcomes the information loss incurred in selecting high frequency categorical attributes in Huang's algorithm. And secondarily, we propose an improved distance function based on Gower's distance for handling mixed data in EDM.

3.0 A PARTITION-BASED FEATURE SELECTION METHOD FOR MIXED DATA-A FILTER APPROACH

The mathematical notations are described in sub-section 3.1. The time complexity analysis is given in sub-section 3.2. In section 3.3, the proposed filter-based feature selection method was discussed. Finally, sub-section 3.4 gives a discussion on cluster validation metrics (CVM).

3.1 Preliminaries

The mathematical notation and the meaning of symbols used in the subsequent paragraphs are presented in Table 3.

Table 3: Mathematical notation and meaning

Notation	Meaning
D_n	A dataset with n features
C	cluster
n	feature
$Freq_{C\|D}(n_i)$	count of n in C in reference to D_i
Instance	observation
gist	summary
Σ	Summation or sigma
|…|	Absolute value
$\in$	Element of
S	Set. A set is a collection of similar objects
U	Universal set
(a, b)	An ordered list or pair of values or a 2-tuple
>	Greater than
{,}	Set brackets
O	Big-O notation. It describes the limiting behavior of a function, when the argument inclines towards a particular value of infinity.

Suppose a dataset D consisting of I objects, where every object has n features (n_{cat} categorical and n_{con} numerical) and $D_n (1 \le n \le 1)$ denotes the *n-th* feature. The numerical features are standardized to median scale.

Definition 1: Given a cluster C with n feature value $n_i \varepsilon D_i$, the frequency of n in C in reference to D_i is defined as:

$$Freq_{C|D}(n_i) = |\{instance \mid instance \hat{I} C, instance.D_i = n_i\}|$$

Definition 2: Given a cluster C, its gist (CG) is given as: $CG = \{p, gist\}$ where p is the cluster density $C(p = |C|)$, gist is summarized as frequency for categorical features and centroid for numerical features:

$$\mathrm{dif}(C_i^{(1)}, C_i^{(2)}) = 1 - \frac{1}{|C_1|.|C_2|}\sum_{p \in C_1} \mathrm{Freq}_{C_1|D_1}(p_i).\mathrm{Freq}_{C_2|D_1}(p_i) = 1 - \frac{1}{|C_1|.|C_2|}\sum_{r \in C_2} \mathrm{Freq}_{C_1|D_1}(r_i).\mathrm{Freq}_{C_2|D_1}(r_i)$$

This dissimilarity measure is predominantly applied to the measurement of dissimilarity between observations in objects. The observation can be numerical or categorical or both. The distance is always a number between 0 (identical) and 1(maximally dissimilar).

3.2 Gower coefficient of similarity

Given an object i and j, the similarity between them is measured by the following equation,

$$S_{ij} = \sum_{k=1}^{n} s_{ijk}\delta_{ijk}w_k \Big/ \sum_{k=1}^{n} \delta_{ijk}w_k \tag{1}$$

where the number of features is denoted by n, s_{ijk} is the similarity between 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 its available for both objects, and w_k is the feature weights. Gower suggested the application of Manhattan distance for numeric features ensuring the numeric features are range (maximum minus minimum values) normalized. For categorical feature, he suggested to assign a value of zero, if the categories match and 1 if the categories do not match.

3.3 Proposed Method- modified Gower coefficient

Suppose a data set D consist of n numerical and p nominal features in a matrix X given as $D = \{m, p\}$. 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. 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, we get the modified equation as:

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

Using (3), we propose 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 leverage 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. On the basis of this, we present our proposed algorithm that is initialized by developing a similarity matrix W from the dataset D. From W, we developed a normalized matrix 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 that builds on the principle of leave-one-out feature elimination over W in measuring the importance of each feature fp_{ij} 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 fp_{ij} The pseudo-code of the proposed method which is termed as Unsupervised Feature Selection approach for Mixed data (UFSM) is presented in Algorithm 1.

Algorithm 1

Algorithm 1: Unsupervised Feature Selection (UFS) approach for Mixed Data Input: X: m x n dataset with m objects and n features $D\{F_1, F_2, ..., F_n\}$ //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. Preprocessing 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_i} c. For each co_i check and resolve issues like skewness, kurtosis, multicollinearity, outliers and missing values. Save result in matrix D_{cont_i} d. Measure association between D_{ca_i} and D_{ca_i} with factor analysis. Save result in M 2. Distance measurement step: a. Pass the matrix M to the modified Gower equation in (3). 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

3.4 Time Complexity Analysis

To assess the time complexity, we begin an initial hypothesis for the final number of CS is p in S $(4 \leq S \leq 6)$ cycles for every categorical feature comprising of unique values of q. We now analyze the time complexity in every step of the proposed method. The first step determines the data type of variable (v) in training set (N) with attribute numbers (a) and classifies it as categorical (a_{cat}) or numerical (a_{cont}). The worst-case scenario for step 1 is $O(v - p(\sum_{i=1}^{a_{cat}a_{cont}} v_i + a_N))$, however, in an optimal case the time complexity is expected to be $O(v.p.a)$. The time complexity of step 2 $O(a^2)$ is due to forward search for determining the clustering tendency. At step 3, the time complexity will be linear at $O(a)$. Thus, the overall algorithm complexity is estimated to be $O(S.v.p(\sum_{i=1}^{a_{cat}a_{cont}} v_i + a_N))$. Also, the complexity of time for each step us linear with dataset size, feature number and total groups. This implies that our algorithm is scalable for high dimensional data.

3.5 Cluster Validation Metric

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 [30]. 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. 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. The relative CVM evaluate the cluster by exercising varied parameter values in an algorithm (e.g., several reiterations of the cluster numbers). In this paper, we have used the Silhouette Coefficient (SC) metric as the CVM. The best value of SC is 1 and the worst value is -1. The SC values near zero indicate 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.

4.0 CASE STUDY

We now present a case study to illustrate our proposed approach. We used a school panel level dataset for academic session 2012-2013 for the state of Delhi, acquired from the District Information System for Education (DISE) [31]. The dataset consisted of six comma separated data files on school demographics (such as school name, location, and address), school facilities, enrollment 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. So, we inner joined these six data files into one. This resulted in 183 features for 5064 schools. The features contained both categorical and numerical data types. Of these 5064 schools there were 1079 kindergarten schools, 1495 primary level (grade 1-grade 5), 525 primary to upper-primary (grade 1-grade 8), 1274 primary to higher secondary (grade 1-grade 12), 650 upper-primary to higher secondary level (grade 6-grade 12), 41 upper-primary level only (grade 6-grade 8). For this study, we are focused on primary level schools because we were interested in analyzing and comparing the factors responsible for student enrollment in primary level schools. For validation purpose, we have deposited the data files on IEEEDataPort[1]

4.1 Data Analysis

The school panel level dataset required preprocessing to simplify the task of knowledge discovery [32]. The experiments were designed and conducted using R programming language. In Fig. 1, we show the demographic distribution of 1495 schools at the primary level in academic session 2011-2012. The Fig. 1A, visualizes 1251 schools located in urban areas. From the total 675 schools are co-educational, 470 schools are boys only and 106 schools are girls only. Similarly, there are 244 schools in rural areas, consisting of 144 co-educational schools, 83 boys' school and 17 girl's school. In Fig. 1B, we have shown the type of school distribution based on its location. A majority of the coeducational schools are located in North-East Delhi (n=187) closely followed by South Delhi (n=138) and East Delhi (n=110). The North-West Delhi has the highest number of boys only schools (n=133). The districts North-East Delhi (n=26) and North-West Delhi (n=24) have the highest number of girls only schools as compared to other districts. In Fig. 1C, we depicted the school type distribution based on whether it was operated by the government or private agencies. A majority of coeducational schools (n=644) are unaided and operated by Municipal Corporation of Delhi (MCD), followed by MCD aided 548 boy's school and 118 girl's school. Continuing further, in Fig. 1D, we reported the enrollment distribution based on instruction medium. It is interesting to notice the coeducational schools impart instruction was in English, whereas the boys and girls school impart instruction were in Hindi. Finally, there are 20 boys, 17 girls and 06 coeducational schools where Urdu language was the instruction medium.

[1] http://dx.doi.org/10.21227/efxf-ck70

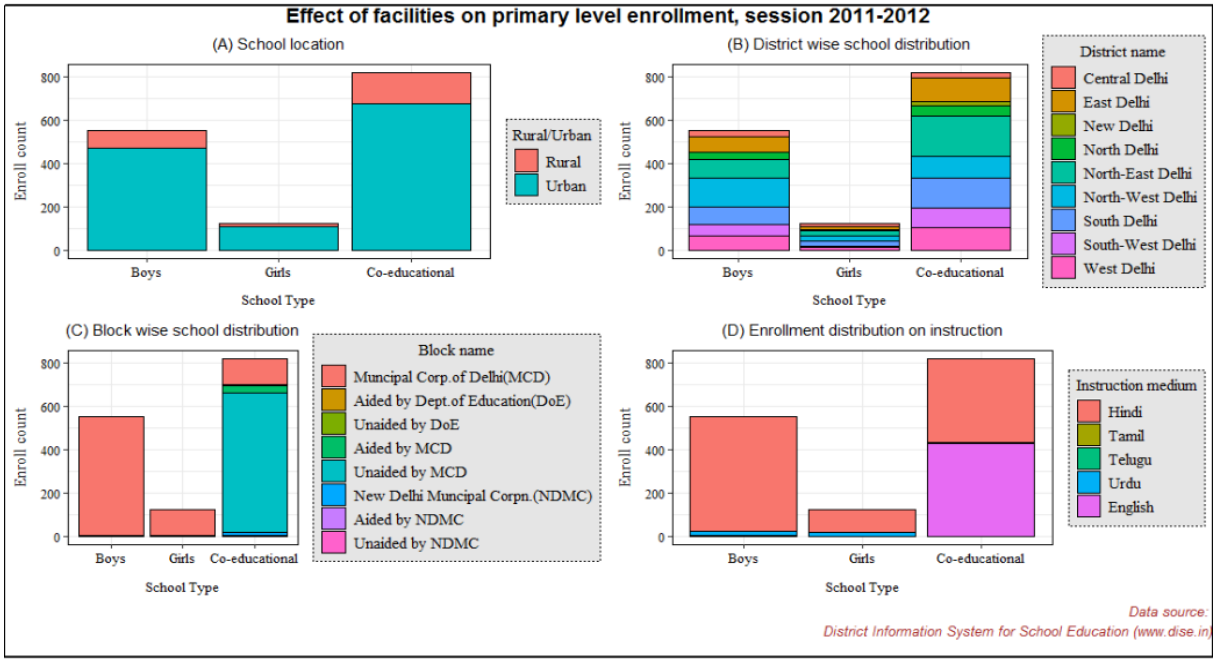

Fig. 1: Primary school enrollment and facilities distribution

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 Fig. 2 denoted as red colored dots are the outliers. From Fig. 2, it is evident the primary schools encompassing only five grade levels with 50 boys' toilets or 60 girl's toilets or above 50 classrooms are clear indications of outliers.

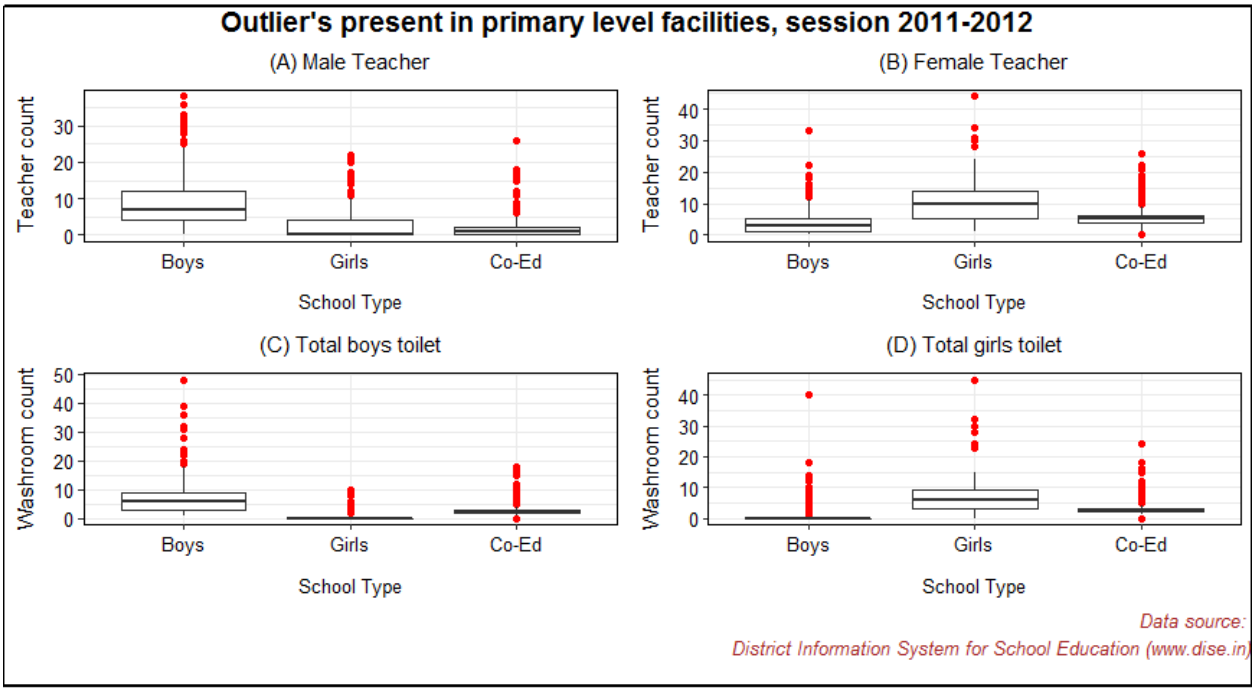

Fig. 2: Outliers in numerical features

For the purpose of our analysis, we performed a partial outlier treatment. Note that, we only considered the features identified in Fig. 2, like `toilet facility for boy and girl`, `total number of classrooms in school` and `blackboards`.

After accounting for partial outlier treatment depicted in Fig. 3, the data dimension reduced to 1,158 primary schools in 51 features. Of these, there were 401 boy's schools, 90 girl's schools, and 874 co-educational schools.

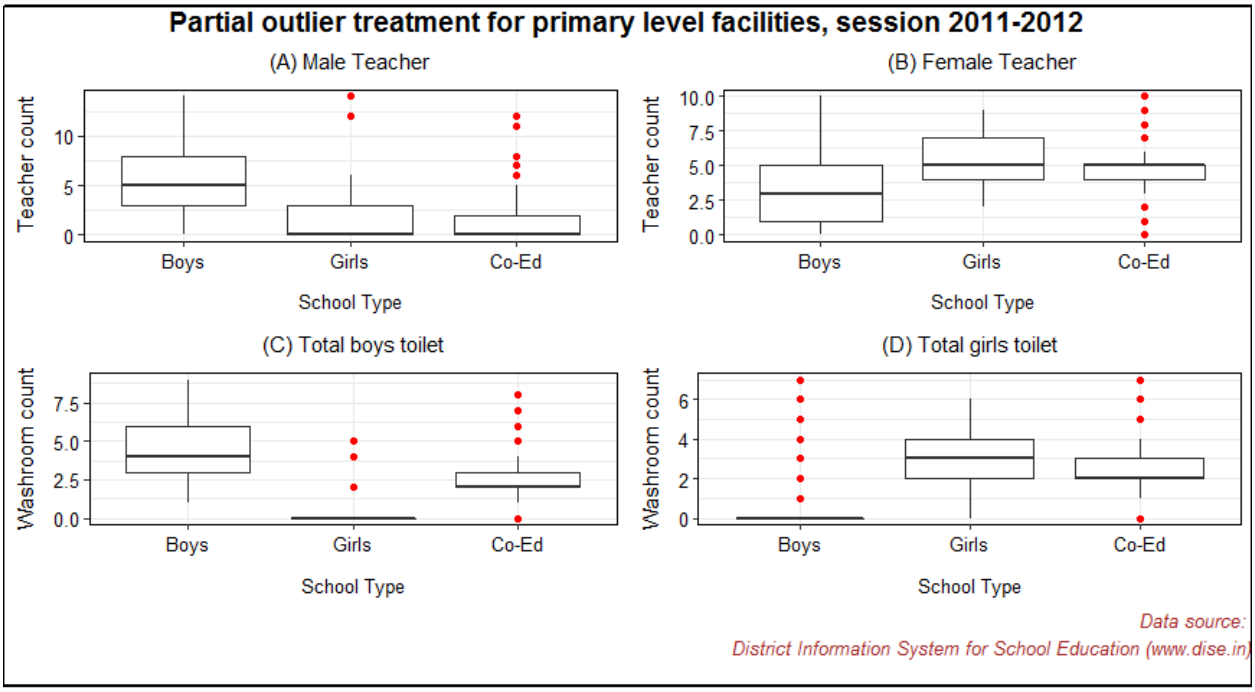

Fig. 3: Partial outlier treatment for numerical features

By this stage, we had reduced the original 183 features to obtain 17 numerical and 6 categorical significant features in 1,158 primary level schools. At this stage, its capability in the simple statistical exploration was somewhat limited due to the multivariate nature of the dataset. The need to group multiple variables of different data types was imminent as this involved a mixed dataset. Thereby, clustering allowed using multiple attributes to identify similar groups in an unsupervised way.

5.0 EXPERIMENTS

In this section we discuss the experimental setup to validate the feasibility of the proposed approach.

5.1 Experimental Setup

We chose 5 mixed datasets (datasets with categorical and numerical features) from UCI Machine Learning repository. Initially, the K-Prototypes algorithm was applied directly to these datasets with a fixed number of four clusters for fair comparison. In all such cases the CVM has returned lower validation scores as shown in Table 4. Thereafter, we illustrate our improved results in applying the UFS based K-Prototypes algorithm approach in Table 5.

Table 4: Application of K-prototypes algorithm on mixed data

S.No.	Dataset name	Attribute Type			Cluster validation metric
		Number of categorical attributes	Number of numerical attributes	Missing values	Silhouette Coefficient
1.	Automobile	10	16	Yes	0.58
2.	Auto mpg	4	5	No	0.55
3.	Census income	9	5	Yes	0.71
4.	Credit approval	10	6	Yes	0.31
5.	Echocardiogram	2	10	Yes	0.56

We now briefly discuss the CVM for various datasets shown in Table 5. Off the 5 datasets in Table 5, we can see that after applying the proposed method, the number of features is reduced (see columns 'Number of categorical attributes', 'Number of numerical attributes'). In comparing the results shown in Table 5 with the results in Table 4, we have obtained a better score for Silhouette Coefficient for most datasets.

Table 5: Application of K-prototypes algorithm using the proposed UFS based approach on mixed data

S.No.	Dataset name	Attribute Type			Cluster validation metric
		Number of categorical attributes	Number of numerical attributes	Missing values	Silhouette Coefficient
1.	Automobile	5	9	No	**0.62**
2.	Auto mpg	2	2	No	**0.60**
3.	Census income	7	2	No	**0.76**
4.	Credit approval	6	4	No	**0.39**
5.	Echocardiogram	1	6	No	**0.60**

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. In Table 6, we show a comparative evaluation of existing partition-based algorithms namely PAM and CLARA with our proposed approach on 05 UCI ML mixed datasets including the school panel level dataset. Since k-means and Fuzzy C-means method work for continuous data only, they were eliminated from the comparison. The k-modes algorithm works only for categorical data. So, it was also eliminated from the comparison. Evidently, our approach outperforms similar partition-based clustering algorithms for mixed data.

Table 6**:** Comparison of the proposed approach with other partition-based algorithms for mixed data

S. No.	Dataset name	Algorithm	Cluster Validation Metric- Silhouette Coefficient
1.	Automobile	PAM	0.71
		CLARA	0.67
		Proposed approach	**0.75**
2.	Auto mpg	PAM	0.78
		CLARA	0.77
		Proposed approach	**0.82**
3.	Census income	PAM	0.69
		CLARA	0.78
		Proposed approach	**0.80**
4.	Credit approval	PAM	0.72
		CLARA	0.81
		Proposed approach	**0.84**
5.	Echocardiogram	PAM	0.74
		CLARA	0.76
		Proposed approach	**0.78**
6.	DISE School panel level dataset	PAM	0.33
		CLARA	0.41
		Proposed approach	**0.56**

5.2 Empirical Results

To remove bias, we split each dataset into 70% training and 30% testing set, using a 10-cross validation measure. Each dataset was tested 10 times by random shuffling to ensure the results were not biased dependent on data sequences. In Table 4 records the original number of features to which the K-prototypes algorithm was applied directly. In Table 5, we show the reduced feature set obtained with proposed feature selection approach. The difference is imminent. It is because we applied a common distance metric for both numerical and categorical

features, before applying the K-prototypes method. To further evaluate the efficiency of the proposed approach, we test it with other partition-based methods for mixed data. We have shown this comparative evaluation in Table 6. Thus, we prove the superiority of our method.

6.0 DISCUSSION

In this section, we provide a discussion on the school panel level dataset because it was the pre-eminent idea of this paper. Initially, we applied the K-Prototypes algorithm directly to the sample of 2581 primary level schools with 183 features. Using the elbow method 2 clusters were chosen as shown in Fig.4. These clusters were tested for purity using silhouette coefficient. The average silhouette coefficient was 0.51 as shown in Table 7.

Table 7: Application of K-Prototypes algorithm on school panel level dataset

Clusters	No preprocessing		Cluster Validation Metric
	Number of categorical features	Number of numerical features	Silhouette Coefficient
2	35	98	0.51

Thereafter, we applied our approach as discussed in sub-section 4.1. We obtained reduced number of features and maximum coherence in 2 clusters. The average silhouette coefficient was 0.56 as shown in Table 8, indicates dense groups.

Table 8: Application of K-Prototypes algorithm using the UFS approach on school panel level dataset

clusters	After preprocessing		Cluster Validation Metric
	Number of categorical features	Number of numerical features	Silhouette Coefficient
2	6	17	0.56

Using the elbow method, we determined the existence of two groups as shown in Fig. 4.

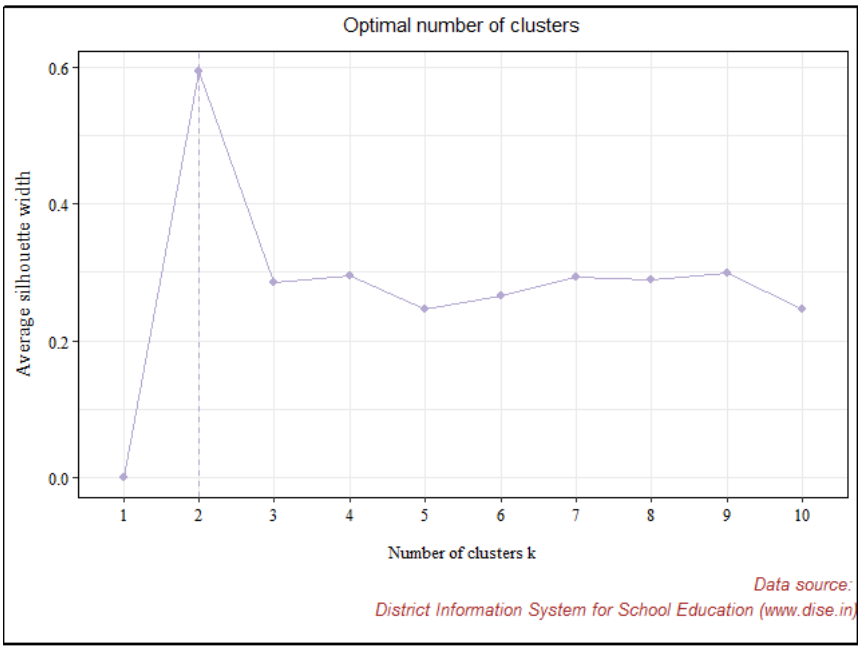

Fig. 4: Optimal number of clusters

We now present a discussion on these 2 clusters. The cluster 1 consist of 499 government schools managed predominantly by Municipal Corporation of Delhi (MCD). Of these, more than 350 schools are boys only with maximum enrollments in grade 5 level. Over 400 schools are operating from government owned buildings. The median number of teachers with professional qualification is surprisingly higher than teachers with graduate qualifications. It is also interesting to note; the number of male teachers is greater as compared to female teachers at the primary level in government schools. And the majority instruction medium was Hindi. Our finding is similar to a recent World Bank study [33] which highlights the appointment of male teachers as heads of primary schools in Rajasthan. This study also highlights the scarcity of female teachers at the primary school levels. There could be similar un-stated practices in other states of India affecting the career opportunities of women or certain specific social groups. We reckon there should be more research on this qualitative aspect that affects equality and inclusion.

The cluster 2 consisted of 659 private unaided schools. A majority of the schools were coeducational with a median number of 5 girl schools. More than 450 schools were operated in rented buildings. While there were no male teachers employed in private unaided schools, there was a minor difference between the median numbers of teacher with either graduate or professional qualifications. This clearly indicates that private primary level schools are not concerned whether or not if the teacher is a graduate. These findings conform with the Right to Education (RTE) act passed by Indian Government in 2010. According to RTE act, there were 7 million teachers employed in private primary schools in India who lacked the basic of a Bachelor of Education (B. Ed) degree [34]. The RTE act has directed the ministry of education to ensure that by year 2019, such teachers employed in private schools acquire the minimum education qualification.

7.0 CONCLUSION AND FUTURE WORK

This study discusses a novel partitional filter-based approach for mixed dataset in EDM. Through extensive experiments we have exhibited the limitation of the existing K-Prototypes algorithm. Our approach leverages the preprocessing techniques as a precursory step. The proposed approach has a linear time complexity that is capable of being scaled to larger datasets. Empirical results on UCI datasets and a real-world educational dataset have proven that the proposed approach elicited similar or better performance. Although we obtained good results in this study, the limitations that exist will be explored in the future work. For instance, the proposed approach was fully dependent on the data distribution.

ACKNOWLEDGMENT

We would like to thank Dr. Rashmi Gangwar (Consultant) at Water Sanitation and Hygiene (WASH) program at UNICEF, who helped us in acquiring the DISE dataset. This work was supported by the University of Malaya research under Grant GPF006D-2019.

REFERENCES

[1] Mollaee, M. and M.H. Moattar, *A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification.* Biocybernetics and Biomedical Engineering, 2016. 36(3): p. 521-529.

[2] Novaković, J., *Toward optimal feature selection using ranking methods and classification algorithms.* Yugoslav Journal of Operations Research, 2016. 21(1).

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

[4] Márquez-Vera, C., et al., *Early dropout prediction using data mining: a case study with high school students.* Expert Systems, 2016. 33(1): p. 107-124.

[5] Asif, R., et al., *Analyzing undergraduate students' performance using educational data mining.* Computers & Education, 2017. 113: p. 177-194.

[6] Dutt, A., M.A.B. Ismail, and T. Herawan, *A Systematic Review on Educational Data Mining.* IEEE Access, 2017. 5: p. 15991-16005.

[7] Saxena, A., et al., *A review of clustering techniques and developments.* Neurocomputing, 2017. 267: p. 664-681.

[8] Huang, Z. *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)*. 1997. Singapore.

[9] Kacem, M.A.B.H., C.-E.B. N'cir, and N. Essoussi. *MapReduce-based k-prototypes clustering method for big data*. in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2015. Paris, France: IEEE.

[10] Kumar, M., K.S. Rani, and C.R. Rao. *Clustering voluminous of heterogeneous data*. in *2017 International Conference on Computer Communication and Informatics (ICCCI)*. 2017. Coimbatore, India: IEEE.

[11] Hotelling, H., *Analysis of a complex of statistical variables into principal components.* Journal of educational psychology, 1933. 24(6): p. 417.

[12] Cai, J., et al., *Feature selection in machine learning: A new perspective.* Neurocomputing, 2018. 300: p. 70-79.

[13] Li, J., et al., *Feature selection: A data perspective.* ACM Computing Surveys (CSUR), 2018. 50(6): p. 94.

[14] Gower, J.C., *A general coefficient of similarity and some of its properties.* Biometrics, 1971: p. 857-871.

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

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

[17] Šulc, Z., M. Matejka, and J. Procházka. *Modifications of the Gower similarity coefficient*. in *19th Conference of Applications of Mathematics and Statistics in Economics-(AMSE 2016)*. 2016. Banská Štiavnica, Slovakia.

[18] Lin, D. *An information-theoretic definition of similarity*. in *15th International Conference on Machine Learning*. 1998. San Francisco: Morgan Kaufmann Publishers.

[19] Huang, Z., *Extensions to the k-means algorithm for clustering large data sets with categorical values.* Data mining and knowledge discovery, 1998. 2(3): p. 283-304.

[20] Foss, A.H., M. Markatou, and B. Ray, *Distance Metrics and Clustering Methods for Mixed-type Data.* International Statistical Review, 2018.

[21] Liu, S. and M. d'Aquin. *Unsupervised learning for understanding student achievement in a distance learning setting*. in *2017 IEEE Global Engineering Education Conference (EDUCON)*. 2017. IEEE.

[22] Dumdumaya, C. and M.M. Rodrigo. *Predicting Task Persistence within a Learning-by-Teaching Environment*. in *26th International Conference on Computers in Education*. 2018. Philippines: Asia-Pacific Society for Computers in Education.

[23] Hashima, A.S., A.K. Hamoud, and W.A. Awadh, *Analyzing students' answers using association rule mining based on feature selection.* Journal of Southwest Jiaotong University, 2018. 53(5).

[24] Zaffar, M., M.A. Hashmani, and K. Savita. *Performance analysis of feature selection algorithm for educational data mining*. in *2017 IEEE Conference on Big Data and Analytics (ICBDA)*. 2017. IEEE.

[25] Velmurugan, T. and C. Anuradha, *Performance evaluation of feature selection algorithms in educational data mining.* Performance Evaluation, 2016. 5(02).

[26] Aldikanji, E. and K. Ajami, *Studying Academic Indicators within Virtual Learning Environment Using Educational Data Mining.* arXiv preprint arXiv:1612.01090, 2016.

[27] Dutt, A. and M.A. Ismail. *Can we predict student learning performance from LMS data? A classification approach*. in *3rd International Conference on Current Issues in Education (ICCIE 2018)*. 2019. Universitas Negeri Yogyakarta, Indonesia: Atlantis Press.

[28] Hamoud, A.K., *Classifying Student's Answers using Clustering Algorithms based on Principal Component Analysis.* Journal of Theoretical & Applied Information Technology, 2018. 96(7).

[29] Amershi, S., C. Conati, and H. Maclaren. *Using feature selection and unsupervised clustering to identify affective expressions in educational games*. in *8th International Conference on Intelligent Tutoring Systems*. 2016.

[30] Zhao, X., J. Liang, and C. Dang, *Clustering ensemble selection for categorical data based on internal validity indices.* Pattern Recognition, 2017. 69: p. 150-168.

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

[32] Bandaru, S., A.H. Ng, and K. Deb, *Data mining methods for knowledge discovery in multi-objective optimization: Part A-Survey.* Expert Systems with Applications, 2017. 70: p. 139-159.

[33] Ramachandran, V., et al., *Getting the Right Teachers Into the Right Schools: Managing India's Teacher Workforce*. 2017: The World Bank.

[34] Sarkar, C.C., *Right of Children to Free and Compulsory Education Act, 2009 and its Implementation*, in *India Infrastructure Report 2012*. 2016, Routledge (Taylor & Francis Group): New Delhi, India. p. 71-81.