Clustering Algorithm for Educational Data Mining: A Systematic Review of Literature and Techniques

M. Durairaj\(^1\). C. Vijitha\(^2\)

\(^1\)Assistant professor School of Computer Science Engineering and application,
\(^2\)Research scholar, School of Computer Science Engineering and application,
\(#1\) Bharathidasan University, Tiruchirappalli, India
\(#2\) durairaj.m@csbdu.in  \(#2\) vijimeena29@gmail.com

Abstract—Educational data mining is a very popular research area for studying the behaviour of students based upon their past performance. As Education is very basic need, which must be given to all, the study of student behaviour plays a vital role. Grouping students on the basis of their performance helps to form a good team for any competitions to represent from an institute or university. Motivational lectures and extra focuses are needed to the students who are having poor performance in academics, which could inspire them to improve their academic and social performances. Some of the data mining techniques and clustering algorithms could be effectively applied on students’ data to retrieve the understandable patterns for performance analysis. This article characterizes the application of clustering algorithm and related processes within the context of a data mining. Also, a systematic review and comparison analysis on the recently published articles on behavioural analysis of students based upon their past academic performance is presented.

Keywords — Datamining, Clustering, K-Means clustering, Students’ behavioural analysis.

I. INTRODUCTION

Data Mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. With the widespread use of databases and the explosive growth in their sizes, organizations are faced with the problem of information overload. The problem of effectively utilizing these massive volumes of data is becoming a major problem for all enterprises. Data mining techniques support automatic exploration of data and attempts to source out patterns and trends in the data and also infers from these patterns which will help the user to support review and examine decisions in some of the related business or scientific area.

Clustering or clustering analysis is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It is the main task of exploratory data mining and a common technique for statistical analysis used in many fields, including machine learning, pattern recognition, and image analysis information retrieval and Bioinformatics.

Azhar Rauf, et. al. [1] proposed a method known as K-mean clustering, it calculates initial centroids instead of random selection, due to which the number of iterations is reduced and elapsed time is improved. Jaideep Vaidya[2] proposed a privacy preserving K-means clustering method over vertically partitioned data when different web sites contain different attributes for a common set of entities. N. Sivaram[3] surveyed the applicability of clustering and classification algorithms for recruitment data mining techniques that fit the problems which are determined. A study has been made by applying K-means, fuzzy C-means clustering and decision tree classification algorithms to the recruitment data of an industry. Md. Hedayetul Islam Shovon[4] presented a paper on prediction of student academic performance by applying K-means clustering algorithm. The student’s evaluation factor like class quizzes, mid and final exam assignment are studied. It is recommended that all these correlated information should be conveyed to the class advisor before the conduction of final exam. This study will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students. Sajadin Sembiring [5] discussed the application of Smooth Support Vector Machine (SSVM) classification and kernel k-means clustering techniques. The results of this study reported a model of student academic performance predictors by employing psychometric factors as variables.
predictors. Oyelade, O. J[6] presented a method of using K-means clustering algorithm for the prediction of Students’ Academic Performance. The ability to monitor the progress of students’ academic performance is a critical issue to the academic community of higher learning. This paper aims to present a systematic review on different clustering techniques applied for educational data mining to predict academic performance of students and its implications. Sunita B. Aher[7] presented a ‘Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data’ for recommend to new student who has recently enrolled for some course. Feng-Hsu Wang[8] presented Effective personalized recommendation based on time-framed navigation clustering and association mining for Personalized recommendation by predicting user-browsing behavior using association-mining technology has gained much attention in web personalization research area. Mofreh A. Hogo[9] discussed about an Evaluation of e-learning systems based on fuzzy clustering models and statistical tools for introduces a hybridization approach of AI techniques and statistical tools to evaluate and adapt the e-learning systems including e-learners.

II. CLUSTERING MODELS AND METHODS

To review the literatures, we determined the research area, expected research goal and research scope. The research area is on analyzing students’ academic performances that applied data mining techniques. The research goal is to design a clustering model and techniques, in the context of data mining, applied to students’ academic performance analysis and to suggest directions for future research. The research scope is the literature on the applications of data mining tools on students’ data to accumulate knowledge published recently. We chose eight articles which applies clustering algorithms for the behavioural analysis of students’ academic performances and analyzed the selected articles to draw some conclusions and identify some future research directions. The following section, we describe clustering models applied and methodologies devised by the authors of selected articles to analyze students’ behavioural data.

2.1 Clustering Models For Behavioural Analysis

As previously mentioned, the different clustering models and corresponding algorithms used for mining data to retrieve students’ academic performances reviewed in this work are as depicted in Table 1.

<table>
<thead>
<tr>
<th>Clustering models</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Connectivity Model</td>
<td>Hierarchical clustering</td>
</tr>
<tr>
<td>2 Centroid Model</td>
<td>K-means algorithm</td>
</tr>
<tr>
<td>3 Distribution Model</td>
<td>---</td>
</tr>
<tr>
<td>4 Density Model</td>
<td>DBSCAN &amp; OPTICS</td>
</tr>
<tr>
<td>5 Subspace Model</td>
<td>Bi-clustering</td>
</tr>
<tr>
<td>6 Group Models</td>
<td>---</td>
</tr>
<tr>
<td>7 Graph Based Models</td>
<td>---</td>
</tr>
</tbody>
</table>

2.1.1 Basic K-mean Clustering Algorithm:

Azhar, et al.[1] stated that according to the basic K-mean clustering algorithm, clusters are fully dependent on the selection of the initial cluster centroids. K data elements are selected as initial centers and then the distances of all data elements are calculated by Euclidean distance formula. Data elements having less distance to centroids are moved to the appropriate clusters. The process is continued until no more changes occur in clusters. The following figure 1 shows steps of the basic K-mean clustering algorithm steps.

Figure 1. Steps of basic K-mean clustering algorithm

The design of the system requires the complete understanding of the problem domain. The data sets and the input attributes are determined through knowledge engineering in an IT industry. The process...
knowledge acquisition is a process that involves defining the problem, identifying relevant stakeholders, and learning about current solutions to the problem. It also involves learning domain-specific terminology, description of the problem and restrictions of it. In this step, interviews were conducted to the domain experts to obtain required information to solve the problem, knowledge extraction was made with the collected information and a knowledge base was built. The knowledge base construction comprises collection of sample data, and deciding which data will be needed in respect to data mining knowledge discovery goals including its format and size.

Knowledge base

Industry Database

Formation of datasets

Eliminate data

Data normalization

Data transformation

Create derived attributes and discrete

Applying clustering techniques

Construct decision tree using ID3, C4.5 and CART

Evaluate the algorithm with highest accuracy

Choosing rules that best fits the problems

Figure 2. Process of data analysis

Information of all the applicants for an industry for the past two years were collected by contacting the HR department. The mining process begins with the step to gather knowledge from the domain experts. Knowledge acquisition is a process that includes elicitation, collection, analysis, modeling and validation of knowledge for knowledge engineering represented in Figure 2. Some of the important issues involved in knowledge acquisition are the knowledge hidden within the domain experts and is not with a single expert. Interviews were conducted with the domain experts to understand the problem and the knowledge required to solve the problem. The knowledge acquired is used along with the recruitment database maintained in the industry to form the dataset for experimentation. The data collected from the industry is complex and have noisy, missing and inconsistent data. The data is preprocessed to improve the quality of data and make it fit for the data mining task. The data used are transformed into appropriate formats to support meaningful analysis. Some more attributes are derived using the acquired knowledge to support the mining process.

Clustering techniques were applied for the data mining and decision trees were constructed with different construction methods such as ID3, C4.5 and CART. The data collected contains both nominal and continuous data; hence ID3 algorithm could be applied only after discretization. Numerical data were discretized to multiple intervals manually. The constructed models were reviewed and evaluated before it is used for decision support. The models were evaluated using accuracy as the criteria to assess the performance of the method. Constructive rules were extracted from the technique; according to the authors which had better accuracy.

2.1.2 Privacy Preserving K-Means Algorithm

Jaideep Vaidhya, et al., [2] formally defined the problem as follows. In clustering algorithm, r be the number of parties, each having different attributes for the same set of entities, n is the number of the common entities. The parties wish to cluster their joint data using the k-means algorithm. The number of clusters required is k. The final result of the k-means clustering algorithm is the value/position of the means of the k clusters, with each side only knowing the means corresponding to their own attributes, and the final assignment of entities to clusters. Each cluster mean be represented as \( \mu_i \), \( i = 1, \ldots, k \). The projection of the mean of cluster i on party j, \( \mu_{ij} \) represent the projection of the mean of cluster i on party j. Thus, the final result for party j is:

- the final value/position of \( \mu_{ij} \), \( i = 1, \ldots, k \)
- cluster assignments: cluster for all points (i = 1, \ldots, n)

The k-means algorithm also requires an initial assignment (approximation) for the values/positions of the k means. This is an important issue, as the choice of initial points determines the final solution. Research has led to mechanisms producing a good initial assignment. Their techniques use classic k-means clustering done over multiple subsamples of the data, followed by clustering the results to get the initial points. For simplicity, we assume that the k means are selected arbitrarily. Since the underlying operations involve k-means clustering, it is quite easy.
to extend our algorithm to search for and start off with good initial means.

Thus, for \( i = 1 \ldots k \), every party selects its share \( \mu_0 \) of any given mean. This value is local to each party and is unknown to the other parties. Require: \( r \) parties, \( k \) clusters, \( n \) points. The privacy preserving \( k \)-means clustering algorithm is as illustrated in Figure 3.

\[
\begin{align*}
1: & \text{for all sites } j = 1 \ldots r \text{ do} \\
2: & \text{for all clusters } i = 1 \ldots k \text{ do} \\
3: & \text{initialize } \mu_{ij}, \text{ arbitrarily} \\
4: & \text{end for} \\
5: & \text{end for} \\
6: \text{repeat} \\
7: & \text{for all } j = 1 \ldots r \text{ do} \\
8: & \text{for } i = 1 \ldots k \text{ do} \\
9: & \mu_{ij} = \mu_0 \\
10: & \text{Cluster}[i] = ; \\
11: & \text{end for} \\
12: & \text{end for} \\
13: & \text{for } i = 1 \ldots n \text{ do} \\
14: & \text{for all } j = 1 \ldots r \text{ do} \\
15: & \{ \text{Compute the distance vector } \sim X_j \text{ (to each cluster) for point } g \} \\
16: & \text{for } i = 1 \ldots k \text{ do} \\
17: & x_{ij} = | \text{data}_{ij} - D \mu_{ij} | \\
18: & \text{end for} \\
19: & \text{end for} \\
20: & \text{Each site puts } g \text{ into Cluster[closest cluster]} \\
21: & \text{closest clusters Algorithm 3} \\
22: & \text{end for} \\
23: & \text{for all } j = 1 \ldots r \text{ do} \\
24: & \mu_{ij} = \text{mean of } j \text{ 's attributes for points inCluster}[i] \\
25: & \text{end for}
\end{align*}
\]

Figure 3. Privacy preserving \( k \)-means clustering algorithm

2.1.3 Fuzzy C-means Clustering

N. Sivaram et al. [3] stated that integration of fuzzy logic with data mining techniques has become one of the key constituents of soft computing in handling the challenges posed by massive collections of natural data [5]. The central idea in fuzzy clustering is the non-unique partitioning of the data into a collection of clusters. The data points are reassigned membership values for each of the clusters and the fuzzy clustering algorithms allow the clusters to grow into their natural shapes [16]. In some cases the membership value may be zero indicating that the data point is not a member of the cluster under consideration. Many crisp clustering techniques have difficulties in handling extreme outliers but fuzzy clustering algorithms tend to give them very small membership degree in surrounding clusters. The non-zero membership values, with a maximum of one, show the degree to which the data represents a cluster. The points at the centre of the cluster have maximum membership values and the membership gradually decreases when one moves away from the cluster centre.

Thus, fuzzy clustering provides a flexible and robust method for handling natural data with vagueness and uncertainty. In fuzzy clustering, each data point will have an associated degree of membership for each cluster. Fuzzy C-means clustering algorithm includes two processes, the calculation of cluster centers and the assignment of points to these centers using a form of Euclidean distance. The process is continued till the cluster center stabilizes. The algorithm incorporates the fuzzy set’s concepts of partial membership and forms overlapping clusters to support it. Each data item is assigned a membership value in the range of 0 to 1 for the clusters. Degree of fuzziness in the clusters is indicated by the parameter named as fuzzification (\( m \)). When the value of \( m \) is equal to 1, the algorithm works like a crisp partitioning algorithm and for larger values overlapping of clusters tends to be more.

2.3.2K-Means Clustering Algorithm

K-means is an old and widely used technique in clustering methods. Here, \( k \)-means is applied to the processed data to get valuable information. The pseudo-code of \( k \)-means clustering is given below.

**Step 1**: Accept the number of clusters to group data into and the dataset to cluster as input values

**Step 2**: Initialize the first \( K \) clusters - Take first \( k \) instances or - Take Random sampling of \( k \) elements

**Step 3**: Calculate the arithmetic means of each cluster formed in the dataset.

**Step 4**: \( k \)-means assigns each record in the dataset to only one of the initial clusters - Each record is assigned to the nearest cluster using a measure of distance (e.g. Euclidean distance).

**Step 5**: \( k \)-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

Classification is data mining task that predicts group memberships for data instances. In educational arena application of the classification method, given works of astudent, one may predicate his/her final grade. The SSVM is further development of Support Vector Machine (SVM). The SSVM generated andsolve an unconstrained smooth reformulation of the SVM for pattern classification using completely arbitrary kernel. SSVM is solved by a very fast Newton-Armijo algorithm and has been extended to non-linear separation surfaces by using nonlinear kernel techniques. The numerical resultsshow that SSVM is faster than other methods and has better generalization ability. Kernel \( k \)-means has been extended to efficient and effective large scale...
clustering [8], since the original KernelK-means had serious problems, such as the high clustering cost due to the repeated calculations of kernel values, or insufficient memory to store the kernel matrix, that make it unsuitable for large corpora. The new clustering scheme is a large scale clustering for Kernel K-means algorithm.

Sajadin Sembiring [4], designed a framework of student performance analysis. Data collection and preparation were carried out on the student data collected from database management system course held at the University Malaysia Pahang (UMP) in third semester of 2007/2008. The questionnaire were used to collect the real data that describing the relationships between behavioral of students (psychometric factors) and their final academic performance. The variable was used in questionnaire are Interest, Study Behavior, Engage Time, Believe, and Family Support.

Sajadin Sembiring [4], designed a framework of student performance analysis. Data collection and preparation were carried out on the student data collected from database management system course held at the University Malaysia Pahang (UMP) in third semester of 2007/2008. The questionnaire were used to collect the real data that describing the relationships between behavioral of students (psychometric factors) and their final academic performance. The variable was used in questionnaire are Interest, Study Behavior, Engage Time, Believe, and Family Support.

Figure 4. Framework of Student Performance Analysis

The number of students was 1000 with three different major in faculty of computer system and software engineering UMP. The sources of collected data were: personal records, academic record of students and course records.

Framework of Student Performance Predictors is as shown in Figure 4. To get better input data for data mining technique, the preprocessing of the data collected was performed. The data was maintained in different tables was joined in a single table. Integration of the data into one files were performed to increase interpretation and comprehensibility. Then, the attributes were discretized to categorical ones. All grades were classified into five groups such as excellent, very good, good, average, and poor. In this step the fields used in the study were determined and transformed when it was necessary. Using normal distribution method, the value of each item in questionnaire was categorized with High, Medium and Low.

Table 2. Sample Student Data Collected

<table>
<thead>
<tr>
<th>CGPA</th>
<th>Interest</th>
<th>Believe</th>
<th>Std.b</th>
<th>Fam.</th>
<th>Eng.</th>
<th>Time</th>
<th>CGPA Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.55</td>
<td>71</td>
<td>71</td>
<td>76</td>
<td>71</td>
<td>76</td>
<td>76</td>
<td>Excellent</td>
</tr>
<tr>
<td>2.59</td>
<td>55</td>
<td>51</td>
<td>56</td>
<td>67</td>
<td>56</td>
<td>56</td>
<td>Average</td>
</tr>
<tr>
<td>2.45</td>
<td>43</td>
<td>43</td>
<td>34</td>
<td>72</td>
<td>34</td>
<td>34</td>
<td>Average</td>
</tr>
<tr>
<td>3.31</td>
<td>71</td>
<td>62</td>
<td>75</td>
<td>67</td>
<td>75</td>
<td>75</td>
<td>Very Good</td>
</tr>
<tr>
<td>2.45</td>
<td>41</td>
<td>32</td>
<td>53</td>
<td>43</td>
<td>53</td>
<td>53</td>
<td>Average</td>
</tr>
<tr>
<td>2.7</td>
<td>52</td>
<td>54</td>
<td>41</td>
<td>65</td>
<td>41</td>
<td>41</td>
<td>Average</td>
</tr>
<tr>
<td>2.6</td>
<td>57</td>
<td>57</td>
<td>53</td>
<td>52</td>
<td>53</td>
<td>53</td>
<td>Good</td>
</tr>
<tr>
<td>2.79</td>
<td>58</td>
<td>58</td>
<td>55</td>
<td>66</td>
<td>55</td>
<td>55</td>
<td>Good</td>
</tr>
<tr>
<td>2.72</td>
<td>54</td>
<td>54</td>
<td>60</td>
<td>31</td>
<td>60</td>
<td>60</td>
<td>Good</td>
</tr>
</tbody>
</table>

The sample students’ data collected were as shown in Table 2. The result of this study indicates that Data Mining Techniques (DMT) capabilities provided effective improving tools for student performance. It showed how useful data mining can be in higher education in particularly to predict the final performance of student. This study expressed the strong correlation between mental condition of student and their final academic performance.

Oyelade, O.J [5] presented a paper on Graded Point Average (GPA) is a commonly used indicator of academic performance. Many Universities set a minimum GPA that should be maintained in order to continue in the degree program. This study provided a simple and qualitative methodology to compare the predictive power of clustering algorithm and the Euclidean distance as a measure of similarity distance.

For example, given a dataset of n data points x1, x2, ... , xn such that each data point is in Rd, the problem of finding the minimum variance clustering of the dataset into k clusters is that of finding k points {mj} (j=1, 2, ..., k) in Rd such that

$$\frac{1}{n} \sum_{i=1}^{n} x_i = 1 \left[ \min_{m_j} d^2(x_i, m_j) \right] \ldots \ldots \ldots (1)$$

is minimized, where d(xi, mj) denotes the Euclidean distance between xi and mj. The points {mj} (j=1, 2, ..., k) are known as cluster centroids. The k-means algorithm provides an easy method to implement approximate solution to Eq.(1). The problem of finding the global minimum is NP-
complete. The flow diagram of cluster algorithm and its functionality is as shown in Figure 6.

Figure 6. Block diagram for implementing the clustering algorithm

Md. Hedayetul Islam Shovon [6] presented a paper on Graded Point Average (GPA) which is a commonly used indicator of academic performance. Many universities set a minimum GPA that should be maintained. Therefore, GPA still remains the most common factor used by the academic planners to evaluate progression in an academic environment. Improving student’s academic performance is not an easy task for the academic community of higher learning. The academic performance of engineering and science students during their first year at university is a turning point in their educational path and usually encroaches on their General Point Average (GPA) in a decisive manner.

Improving student’s academic performance can be immensely useful in “MOOC (Massively Open Online Courses)”. Bisecting Medoids Algorithm) to cluster users based on the time-framed navigation sessions.

Sunita B. Aher[7] presented a combination of clustering technique – simple k-means and association rule algorithm – Apriori to find the result. These results were compared with the results of open source data mining tool Weka. Framework for course recommendation system in e-learning using clustering and association rule mining is as shown in Figure 7. The result obtained using combined approach matches with real world interdependencies among the courses. Other combinations of clustering and association rule algorithms are also discussed here to select the best combination. It was concluded that the Course Recommendation System could help in building intelligent recommender system. This approach of recommending courses to new students can be immensely be useful in “MOOC (Massively Open Online Courses)”. It was concluded that the Course Recommendation System could help in building intelligent recommender system. This approach of recommending courses to new students can be immensely be useful in “MOOC (Massively Open Online Courses)”.

Feng-Hsu Wang [8] proposed a new clustering method, called HBM (HierarchicalRecommendation method to an e-learning web site was presented, including plans of recommendation policies and proposal of new efficiency measures. The effectiveness of the recommendation methods, with and without time-framed user clustering, were investigated and compared. The results showed that the recommendation model built with user clustering by time-framed navigation sessions improves the recommendation services effectively.
known as k smalls. K-means produce tighter clusters than hierarchical clustering, especially if the clusters are globular. K-means clustering technique is fast, robust and easier to understand. It is relatively efficient and it gives best result when data set are distinct or well separated from each other.

As a negative side, it is difficult to predict K-value. With global cluster, it is not worked well. Different initial partitions resulted in different final clusters. It does not work well clusters of different size and different density. The learning algorithm requires apriori specification of the number of cluster centers. Use in the Exclusive Assignment,k-means could not resolve that there are two clusters if there are two highly overlapping data. Learning algorithm is not invariant to non-linear transformations i.e. with different representation of data. Data represented in form of cartesian co-ordinates and polar co-ordinates give two different results. Euclidean distance measures could unequally weight underlying factors. The learning algorithm provides the local optima of the squared error function. Randomly choosing of the cluster center cannot lead us to the fruitful result. It is applicable only when mean is defined i.e. fails for categorical data. It shows inability to handle noisy data and outliers. The algorithm fails for the non-linear data set.

IV. PROPOSED FRAMEWORK FOR STUDENTS’ PERFORMANCE ANALYSIS

We analysed recent data mining applications for students’ performance analysis using different methodologies and techniques. From this study, we proposed a framework for our future work for an efficient evaluation system for adapting e-learning methodologies. The Figure 9 shows different steps involved in the proposed system which could evaluate and adopt e-learning methodologies.
V. CONCLUSION AND FUTURE WORK

In this study, we discussed different methodologies which use marks as a measure for clustering the students. There are many other behaviors by which we can cluster the students. Some of the measures are based upon practical knowledge, class behavior, and talent in particular field, family background. In many applications, authors did not consider any of these measures as it could be very difficult to understand each student. We can enhance the clustering techniques based upon the above mentioned criteria.

From this analysis, we could see that k-means clustering algorithm and deterministic model was used to evaluate the performance of students in higher institutions in many of the applications. It also enhanced the decision making by academic planners to monitor the candidates’ performance in semester by semester by improving on the future academic results in the subsequence academic session. The students’ academic performance could be easily analyzed by the K-means clustering algorithm.

REFERENCES