Performance Evaluation and Comparison of Clustering Algorithms used in Educational Data Mining

N.Valarmathy, S.Krishnaveni

Abstract:*Data mining in educational system has received great interest and has become a new emerging research nowadays. Recently all universities and colleges are generating huge volume of data by conducting online exams and storing lot of information for future purposes. These massive amount of data stored needs some data mining techniques to retrieve some useful and meaningful information from the dataset. The real victory can be achieved only when task specialized is applied so that it can be effective in that area. This paper surveys the application of data mining to traditional educational systems, various well known clustering algorithms, its applications, advantages and disadvantages. This paper also focuses on performance evaluation of some clustering algorithms using educational dataset.*

Index terms: *Clustering algorithms ; Applications; Data mining; performance evaluation of algorithms; education system;*

I. INTRODUCTION

The international consortium has defined educational data mining (EDM) as “an upcoming discipline which concerns about the different techniques for finding out the unique types of data based on educational settings and applying those methods to understand and learn about the behaviour of the students[1]. The educational institutions are producing huge amount of data which can provide clear and deep insights into learners behaviour and their process adopted. A careful processing technique is essential to convert those data into reasonable insights. The EDM process is concerned with developing new methods for exploring data based on the educational settings and using those methods to understand the student’s behaviour and the improvements to be made in the discipline they learn.

As the number of educational institutions is growing rapidly, each institution has to apply some DM techniques which can help them to analyse the student’s performance based on the recorded data and improve them in overcome attrition rates. The various tasks performed by EDM are analysing data, intra-connecting systems for developing new models, for enhancing institutional effectiveness and learning experience. The application of data mining techniques in the context of EDM is limited, but now-a-days various methods have been proposed, already available algorithms were applied for the improving the learning experience and show improvement in their educational institution [2]. EDM must be different from standard DM because of non-independent and multi level hierarchy of educational data [1]. Various tools like python, rapid miner, R-language, Weka which are used in DM are nowadays applied for the improvement of institutional and students’ growth. Predictive modelling technique helps to identify and analyze student’s retention efforts. The novice user or the system administrator cannot make use of the algorithm directly hence it is necessary for a platform to be developed which integrates some combined DM algorithms to analyse these data and produce necessary reports. [45]

Educational data mining and learning analytics have the ability to make available inaction able and visible data into actionable data. Several recommendations are made in this area by the researchers to take necessary actions and help the educators for further improvement. The EDM research must focus on the following ideas like:

- Involving IT departments to collect the data in an effective and usable manner
- Develop a system for using the collected data in instructional decision making process [3].
- The anonymous data collected must be aligned properly across different systems for effective processing [4].
- The researchers must conduct research on usability and effectiveness of data displays [47].
- The suggestions and ideas given by the researchers through displays must help the instructors to be more effective in the classroom with more real-time and data based decision support tools and services.
- This research must also help to identify the students performance from the student information where it will mostly aid them to progress in their studies [5].
- The research must also concentrate in repurposing the developed predictive models from one context to another [15].
- The research must help to align state policies with necessary technical requirements in different learning environments like online, smart class, class room learning etc[21,51].

Revised Manuscript Received on December 22, 2018

N.Valarmathy Ph.D Research Scholar & Assistant Professor, Department of Computer Science (PG & Research), Pioneer College of Arts and Science, Coimbatore, Tamil Nadu, India

Hence the research on EDM must develop a best tool that can help students, instructors and educational institution in their further growth and upliftment. A strong collaboration must be made between the researcher, commercial companies and educational sectors to operate effectively on the data and fast development cycles. Effective partnership can help them to produce a tool which can satisfy all their problems [11].

This paper is compiled as follows. The review of previous work done with educational data mining is summarized in Section II; Section III discusses the various clustering algorithms / techniques applied to educational dataset. Section IV discusses about the advantages, disadvantages and application of different clustering algorithms. Section V provides dataset used in this experimental setup, Section VI shows results and discussion and finally Section VII shows conclusion and future work.

II. Literature review

About 100 papers related to clustering and educational data mining were studied and some of the most techniques and algorithms used by them already is tabulated in the table below:

S.No	**Objective of the paper**	**Algorithms used**	**Database used**	**Authors**
1	A comparative Analysis of clustering Algorithms	K-means, Hierarchical Clustering algorithms	Iris, Haber man, Wine from UCI repository	Pallavi, G Sunila (IJERA)[8]
2	Performance Comparison of Various clustering Algorithms	Simple k-Means, Enhanced K-means, Farthest first, Make density based, Filtered	Abalone and Letter image from UCI repository	Revathi et.al. (IJARCSSE) [6]
3	Comparative Investigation of K-Means and K-Medoid Algorithm of Iris Data	K-Means, K-Medoid	Iris Data Set from UCI Machine Learning Repository	Tiwari and Singh,(IJERD)[46]
4	Evaluate the performance of under graduate students	Combination of ANN and farthest first	Student data of computer science department (NUDM)	S. Chen and X. Liu [49]
5	To predict students behavior in future	UCAM algorithm	Student dataset	F. Getúlio, R. De Janeiro et.al[16]
6	Clustering high dimensional data set using fully automated algorithm	Two phase clustering algorithm	Real life dataset	M. I, D. A. Kashy et.al.[18]
7	Deals with clustering student access patterns	Fuzzy sets and transitive closure	No information regarding dataset	C. Romero, S. Ventura, and E. García [17]
8	To identify the significant variables that affect the performance of under graduate students	c-means clustering	Academic dataset from IUS	M. Pechenizkiy, et.al,[14]
9	To compare the emotional intelligence of students	K-Means clustering	Student dataset	T. Etchells et.al, [20]
10	How to teach students from rural background	K-means clustering	Survey dataset	D. Ibrahim and Zaidah, Rusli [19]
11	To map out the approaches to teaching profiles of teachers in higher education on the basis of ATI scores	Hierarchical cluster analysis	Questionnaire data and interviews	P. Golding and O. Donaldson [50]
12	A new employee profiling system has been created and used in higher educational environment	Employee profiling software	Higher education student dataset	Archer, Elizabeth,et.al, 2014 [32]
13	To compare the emotional intelligence of students	K-Means clustering	Student dataset	T. Etchells, À. Nebot, A. Vellido, P. Lisboa, and F. Mugica [20]
14	Graphically representing institutional growth prognosis and student's progress analysis	Naïve Bayes algorithm	Student dataset	Saranya. S.R. Ayyappan and N. Kumar 2014[27]
15	Analysis of professional dataset during a training camp conducted by a consultancy company.	Clustering technique	Online training data	Hicheur cairns, Awatef et.al,2014[28]
16	To predict the potentiality of students performance who can fail during an online curriculum in LMS	EM, HCA, simple k-means, X-means	Real life dataset provided by juris campus	P. Moreno - Clari, M. Arevalillo-Herraez, and V. Cerveron-Lleo[12]
17	Comparison the various clustering algorithms of Weka	DB Scan, EM, Cobweb, Optics, Farthest First, Simple K-Means	ISBSG and PROMISE repository	Sharma N et al. (IJETAE)[48]
18	Performance of Computer Engineering	DB Scan, EM, Hierarchical, K-means	Zoo, Labor, super market from UCI Machine Learning repository	Tiwari et al. (IOSRJCE)[7]

19	The applications of various DM techniques on evaluating student performance	APIRORI algorithm	Student academic record file	H. Grob, F. Bensberg, and F. Kaderali [28]
20	Identifying the set of weak students based on graduation and post-graduation marks	Association analysis algorithm	Survey dataset	Arora, Rakesh and Dharmendra Badal,2014 [29]
21	It helps to Analyze the performance of low academic achievers in higher education	Bayesian classification method	questionnaire	Sukanya, et.al, 2012 [30]
22	Estimated success chances of curricula by implementing student profiling with story board system	Decision tree	Student dataset from UCI	Sakurai et.al, 2011[31]

III. CLUSTERING TECHNIQUES

Clustering or data grouping is an unsupervised learning task in which finite set of categories are identified as clusters based on the intra class similarity present in the data. Between each cluster there is maximum intra class similarity and minimum inter class similarity[36]. Thus clustering technique is applied to identify the intrinsic grouping between a set of data which are not labelled or grouped [37]. This technique can be used when the classes are not known and it does not analyse class labelled instances as used in classification. The attribute which provides the good similarity must be identified to increase the similarity metric between clusters. Cluster properties can be analysed to identify the profiles which distinguish one cluster from the other. The performance of good clustering technique is measured by its ability to identify the patterns that are hidden and produce maximum intra class similarity and decrease interclass similarity between other objects among clusters [42].

The techniques adopted in clustering are categorized into three methods namely Hierarchical methods, Partitioning methods, and Density-based methods [33]. The partitioning method can be used to determine k clusters which can optimize each clusters based on distance function. Hierarchical methods create a dendrogram by decomposition of database. Density-based methods finds out the dense regions available in the data space which are separated from each another by low density noise regions.

Hierarchical algorithms

Hierarchical methods creates a nested sequence of clusters, using a single or all of inclusive cluster at the top and singleton clusters from the bottom of individual points[38]. The hierarchy created either through top-down (divisive) or bottom-up (agglomerative) manner and is not extended till the extremes. When the desired number of clusters has been obtained the process of merging or splitting is stopped. The method of merging or splitting a pair of cluster based on some criterion is done in each iteration through which proximity between clusters can be increased. The main drawback faced in this approach is that once merging or splitting is done it is not reversible. There are many algorithms developed using this hierarchical technique, some of them are listed below:

CURE

Clustering Using Representatives (CURE) [9] is a bottom-up method using two new ideas. First, clusters are formed using fixed number of well scattered points not involving centroid.

Second a constant factor is used to shrink the representatives towards the clusters and the closest representatives to the clusters are merged together. The technique of using multiple representatives allows clusters to be of various sizes and shrinking method helps to remove the noise and outliers. This technique makes use of partitioning and random sampling to increase the scalability rate.

CHAMELEON

CHAMELEON method is used to increase the quality of clustering which does merging of clusters in an elaborate manner when compared to CURE [38]. In first phase the k-nearest neighbours and graph showing the link between every point are identified. Then algorithm to perform partitioning is recursively used to split the graph points slitted in many small unconnected sub graphs. In phase two, the unconnected sub graphs are considered as the sub cluster and bottom up approach is to merge to combine similar clusters. The merging is performed only when it has similar inter connectivity and closeness between two clusters. Using arbitrary shaped cluster having different density and bottom up merging makes this algorithm more effective than CURE. The performance of this algorithm is high even when the size of database is high but its computational cost increase with its size.

BIRCH

Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH) applies a new concept in framing the hierarchical data structure that is the CF-tree, to compress the data into large number of small sub-clusters and then applies clustering using the summaries instead of using the raw data. Compact summaries are used to represent each sub clusters called as cluster feature (CF) and are stored in the leaf node [39]. The sum of CF of each child is stored in non-leaf nodes. The CF tree is built incrementally and dynamically by inserting an object available in the closest leaf entry. The maximum number of children in the non leaf node and the diameter between two clusters are controlled using two parameters. A new structure can be created and partitioning algorithm is applied to build the CF tree and stored in the main memory. This algorithm works faster but it's highly sensitive to the data order and non spherical clusters of different size.

COBWEB

COBWEB makes uses of the concept of incremental system in hierarchical clustering [38]. The hierarchical tree is formed by incrementally arranging each observation. Each nodule represents a class using probabilistic approach and summarizes object classification and the value corresponding to each attribute in the nodule [10]. This tree formation helps to identify the misplaced attributes and values of the new object. This technique uses the inquisitive estimation measure also called as class efficiency to observe the construction of the tree and its efficiency.

Advantages	Disadvantages
COBWEB performs bidirectional search by using splitting and merging technique.	It is based on the assumption that probability distributions on separate attributes are statistically independent of one another.
COBWEB uses a heuristic evaluation measure called category utility to get highest category utility	
In this algorithm splitting and merging of classes is performed on category utility.	It is quite expensive to store and update the clusters as they are represented using probability distribution.

Density-Based Clustering Algorithms

Density-based clustering algorithms works by grouping its neighboring objects to form clusters using local density function and does not use proximity determined between objects to form clusters[40]. In these methods clusters are considered as dense regions which are separated by low noisy regions. These methods are tolerant to noise and can identify many non convex clusters. This technique effects clustering tendency when there is inherent scarcity of feature space available in high dimensional spaces. The various forms of these algorithms are listed below:

DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) works by finding out the core objects which has least Minpts in the neighborhood [41]. The skeletons of a cluster are formed using the core objects which has overlapping neighborhood and other objects are considered as noise. This algorithm can be applied when it needs to remove the outliers correctly and the order of the input data is more complex. This algorithm is very sensitive to the input parameters and Minpts and breaks down the high dimensional spaces available in the dataset. The advantage and disadvantages of this algorithm is tabulated below:

Advantages	Disadvantages
DBSCAN can find arbitrarily shaped clusters using MinPts parameter	DBSCAN cannot perform well with large differences in densities
The ordering of the points in the database is insensitive and it requires two parameters.	DBSCAN can only result in a good clustering [39] as good as its distance measure is in the function region Query (P).Euclidean distance measure is default.
This algorithm does not need to know the number of cluster in advance.	
DBSCAN has a notion of noise.	

DENCLUE

Density-based Clustering (DENCLUE) algorithm illustrate the force of a point on its neighborhood by making use of an influence function and the overall density of the data space is measured using the sum of influence functions from all the data. Clusters are identified using local maxima of the overall density function and the density attractors. The sum of influence functions of each grid structure is computed and applied. DENCLUE can be used to find arbitrary-shaped clusters which is resistant to noise and insensitive to data ordering.

Partitioning Clustering Algorithms

Partitioning clustering works by decomposing a set of N objects into k clusters so that each partition can be optimized using a predefined criterion function. The center of gravity or the centroid is used to represent each cluster. A relocation scheme is iteratively is typically used to reassign the points between cluster and usually k-number of seeds are randomly selected, and it is optimized using the criterion. Generally sum of squared Euclidean distances or minimization of the square error criterion is used to get the optimized clustering results. Some of the algorithms which use this concept are listed below:

Fuzzy K-Means Clustering

Fuzzy k-means clustering algorithm makes use of a degree of points which belong to one just one cluster alone [26]. Thus this points on the edge of each cluster has lesser degree than those in the center of the cluster. The point in each cluster has an coefficient identified by its degree available within k clusters. The steps followed in this algorithm are listed below:

The fuzzy k-means algorithm [43] is very similar to the k-means algorithm:

- Choose the number of clusters.
- Assign randomly to each point coefficients for being in the clusters.
- Repeat until the algorithm has converged (that is , the coefficients' change between two iterations is no more than i, the given sensitivity threshold)

This algorithm can be used to minimize intra-cluster variance and to have minimum number of local points which depend on the choice of weights initially used [22].

Expectation-Maximization

The expectation maximization algorithm is an advanced statistical technique which makes use of partial membership in classes. It also has enhanced convergence properties hence it is generally preferred than fuzzy-k-means algorithm. This has the following advantages and disadvantages listed below:

Advantages	Disadvantages
• Gives extremely useful result for the real world data set. • This algorithm is used when the size of the dataset is small or region-of interest obtained is not satisfied.	• Algorithm is highly complex in nature.

K-Means

K-means is the most popular clustering algorithm commonly used in all metric spaces. Initially the selection of k cluster centroids are done at random [34, 35]; It then reassigns all the points to its nearest centroids and recomputes centroids for the newly assembled groups. This iterative relocation is performed until the criterion function converges with each other. K-means is highly sensitive to outliers and noise as some are affected by the centroids. The main advantage of this algorithm is it minimizes an objective function when squared error function is used. This algorithm works as follows:

The algorithm steps are

Choose the number of clusters, k.

Randomly generate k clusters and determine the cluster centers, or directly generate k random points as cluster centers.

Assign each point to the nearest cluster center.

Re compute the new cluster centers.

Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed)

This algorithm has the following advantages, disadvantages and applications:

Advantages	Disadvantages	Application
1. Most commonly used and easily implemented 2. Computationally faster method 3. Scalable 4. Faster for low dimensional data 5. Produces tight clusters 6. Find more sub-cluster if data large 7. Cluster number is specified. 8. Work only for well-shaped clusters 9. Fixed number of clusters can make 10. It difficult to predict what K should be.	1. Difficulty in comparing quality of the clusters produced. 2. Fixed number of clusters can make it difficult to predict the value of K. 3. This algorithm does not work well with non-globular clusters. 4. Different initial partitions can result in different final clusters. 5. Not handle non-globular data of different size and densities. 6. not able to identify outliers & noise 7. Restricted to data which has the notion of centre (centroid)	I. Geostatic II. Computation vision II. Market segmentation V. Earth quake study V. Land use

Farthest First

Farthest First algorithm is an adjustment form of K- Means algorithm that places every cluster in the center at the point beyond most of the accessible cluster center. This point must lie inside the data area. Hence this algorithm highly increases the speed in different situations where less reassignment and modification is essential.

The best options to be used in this algorithm are: N and S., where N specifies the number of clusters to be generated and S- denotes the number of random seeds.

K-MEDOIDS

K-Medoids also called Partitioning around Medoids (PAM) technique which creates cluster by making use of medoid which are most centrally available in the cluster. The obtained medoids are highly resistant to outliers and noise. This algorithm begins by randomly selecting an object as medoid from each of the k number of cluster. Then the non selected objects are grouped to the medoid to which it is most similar. It iteratively replaces the non-medoid objects by the medoids which yields the improvement in the cost function. It is an expensive algorithm. It cost more to compare each medoid with the entire dataset in each iteration.

Clustering Large Applications Based On Randomized Search – CLARANS

CLARANS makes use of sampling techniques along with PAM. The clustering process begins by searching a graph where each node gives a possible solution, which is a set of k-medoids. The obtained result after replacing a medoid is known as the neighbor of the current clustering. This CLARANS works by [13] selecting a node and comparing it to a user-defined number of neighbors searching for a local minimum. The process is repeated again from the beginning if there is no perfect neighbor found; otherwise the current clustering is considered as a local optimum. When the local optimum is identified, it starts by randomly selecting a new node and searches for a new local optimum again.

Filtered Cluster

The Filtered clustering algorithm works by filtering the information, pattern or data available in the given dataset [38].

Here the user supplies keywords or a set of samples which contain some relevant information. Each and every new information that is identified, are then compared with the available filtering profile and the information which matches to the keywords is given to the user. Filtering profile should be corrected and verified by the user by giving appropriate feedback on the displayed or given information. This algorithm starts by storing the data points in a kd-tree. In each stage the closest center to every data point is calculated and every center is moved towards the centroid of the connected neighbors. The data for every node are filtered again as they are propagated to the children nodes.

Make Density Based Clustering Algorithm

The make density based clustering algorithm works by using internal wrapping technique. Both the distribution value and density are returned as output. This algorithm can be used when there is uneven number of clusters. In this clustering is performed based on the density of data point in a region. Each cluster in the neighborhood of the given radius (EPS) will have at least minimum number of instances (Min Pts) so that this can be used even when the data has noise and outliers. The data points having same density are connected to form clusters of low density and high density regions [44].

IV. Advantages, Disadvantages And Application Of Clustering Algorithms

The advantages, disadvantages and the application of the three basic clustering algorithms are summarized below

S.No	Algorithms	Advantages	Disadvantages	Application
1	Hierarchical Algorithms	• Embedded flexibility regarding the level of granularity. • It can be used for problems which involves point linkages	• Inability to make corrections once the splitting/merging decision is made. • Lack of interpretability regarding the cluster Descriptors. • Vagueness of termination Criterion. • It is very expensive for huge datasets. • Severe effectiveness degradation in high dimensional spaces due to the curse of dimensionality phenomenon.	• Pattern recognition • Image segmentation • Wireless sensors Networks • City planning • Spatial data analysis
2	Density-Based Clustering Algorithms	• It helps to discover clusters of different size. • Resistance to Noise and outliers	• High sensitivity to the setting of input parameters • Poor cluster descriptors • Unsuitable for high-dimensional datasets because of the curse of dimensionality Phenomenon.	• Scientific literature • Images of satellite • Crystallography of x-ray • Anomaly detection in temperate data

3	Partitioning Clustering Algorithms	• Relatively scalable and simple. • Suitable for datasets with compact spherical clusters that are well-separated.	• High dimensional spaces is ill-defined • Poor cluster descriptors • Reliance on the user to specify the number of clusters in advance • High sensitivity to initialization phase, noise and outliers • Frequent entrapments into local optima • Inability to deal with non-convex clusters of varying size and density.	• Geostatic • Computation vision • Market segmentation • Earth quake study • Land use
4	Density Based Algorithm	• Useful when clusters are not normal • Return both distribution and density • Used when data has noise • Used when outliers in the data • Gives result close to K-mean Algorithms.	• Datasets with altering densities are tricky. • Sensitive to clustering • Parameters Min Points and EPS. • Sampling affects density measures.	• Scientific literature • Images of satellite • Crystallography of x-ray • Geostatic • Earthquake study

V. DESCRIPTION OF DATASET USED

The data set used in this research is either real world data obtained from UCI machine learning repository and widely accepted data set available. For experimental purpose two datasets containing continuous attributes (nominal type) that is all these datasets have the following attributes:

#Attributes for both student .mat.csv(Math course) and student.por.csv (Portuguese language course) datasets:

1. School – student’s school (binary: “GP” – Gabriel Pereira or
“MS” – Mousinho da Silveira)
2. Sex – student’s sex (binary: “F” – female or “M” – male)
3. Age – student’s age (numeric from 15 to 22)
4. Address – student’s home address type (binary:”U”- urban or
“R” – rural)
5. Famsize – family size (binary: “LE3” – less or equal to 3 or
“GT3” – greater than 3)

I. PStatus – parent’s cohabitation status (binary: “T” – living
together or “A” – apart)
II. Medu – mother’s education (numeric: 0-none, 1 – primary
education (4th grade), 2- 5th to 9th Grade, 3- secondary education or 4- higher education)
III. Fedu – Fathers education (numeric 0 – none, 1 – primary
education (4th grade), 2- 5th to 9th Grade, 3- secondary education or 4- higher education)
IV. Mjob – mother’s job (nominal: “teacher”, “health” care related, civil “services” (e.g. administrative or police),”at_home” or “other”
V. Fjob – father’s job (nominal: “teacher”, “health” care related, civil “services” (e.g. administrative or police),”at_home” or “other”
VI. Reason – reason to choose this school (nominal: close to “home”, school “reputation”, ”Course “preference or “other”)
VII. Guardian – student’s guardian (nominal: “mother”, “father
“or “other”)
VIII. Travel time – home to school travel time (numeric: 1<15 min., 2 – 15 to 30 min., 3 – 30 min to 1 hour, 4- > 1 hour)
IX. Study time – weekly study time (numeric: 1 - < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours, Or 4 – >10 hours)
X. Failures – number of past class failures (numeric: n if 1<=n<3, else 4)
XI. Schoolups – extra educational support (binary: yes or no)
XII. Famsup – family educational support (binary: yes or no)
XIII. Paid – extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
XIV. Activities – extra curricular activities (binary: yes or no)
XV. Nursery – attended nursery school (binary: yes or no)
XVI. Higher – wants to take higher education (binary: yes or no)
XVII. Internet – internet access at home (binary: yes or no)
XVIII. Romantic - with a romantic relationship (binary: yes or no)
XIX. Famrel – quality of family relationships (numeric: from 1 –
very bad to 5 – excellent)
XX. Freetime – free time after school (numeric: from 1 – very low to 5 – very high)
XXI. Goout – going out with friends (numeric: from 1 – very low to 5 – very high)
XXII. Dalc – workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
XXIII. Walc - weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)

XXIV. Health – current health status (numeric: from 1 – very bad to 5 – very good)

30. Absences – number of school absences (numeric: from 0 to 93)

#these grades are related with the course subject, Math or Portuguese

31. G1 – first period grade (numeric: from 0 to 20)
32. G2 – second period grade (numeric: from 0 to 20)
33. G3 – final grade (numeric: from 0 to 20, output target)

To verify the performance of the selected 8 algorithms in our research, the student's higher secondary data is being used from UCI machine learning repository. The eight algorithms which has been discussed above have been implemented in JDK and python in order to measure the performance using several metrics and parameters using our students performance analysis datasets. Table below depicts the numbers of instances and attributes of the used datasets.

VI. VI.INTERPRETATION AND RESULTS

S.No	Clustering Algorithm	Attributes	Instances	Clustered Instances	Time taken to build the cluster (Sec)	Sum of squared errors	Number of iterations performed
1	Expectation maximization algorithm Log likelihood value = -9.7878	32	8124	14	8613.74	19	23
2	CLOPE	32	8124	23	6.27	25	20
3	**DBSCAN Epsilon=0.9 Minpts=6**	32	8124	8124	112.24	12	15
4	Filtered cluster	32	8124	2	1.14	20	11
5	Farthest first	32	8124	2	0.06	17	11
6	COWEB (splits=87;merges= 90)	32	8124	172	0.92	16	14
7	K-Means clustering	32	8124	K is defined	4.567	19	23
8.	CLARA	32	8124	1200	0.75	18	20

The above mentioned algorithms has also been compared in terms other evaluation metrics like accuracy, precision and F1-measure which is calculated using the formula given below:

Accuracy - Accuracy is the most widely used performance measure and it calculated as a ratio of number of correctly predicted observation to the total observations.

Precision - Precision is defined as the ratio of number of correctly predicted positive observations to the total number of predicted positive observations.

Recall (Sensitivity) - Recall is defined as the ratio of number of correctly predicted positive observations to the all observations in actual class.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

The above mentioned algorithms were also compared using these 3 metrics using the values obtained from the confusion matrix. The table below shows the obtained results.

$$Ac = \frac{TP + TN}{TP + FP + FN + TN}$$

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. The F1 Score is calculated with precision and recall. It is also called the F Score or the F Measure.

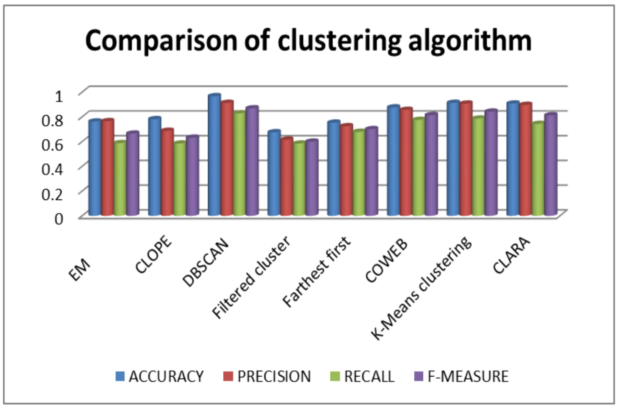
F measure = 2 * { (precision * recall) | (precision + recall)}

S.No	Clustering Algorithm	Accuracy	Precision	Recall	F-Measure
1	Expectation maximization(EM)	0.7623	0.7654	0.5876	0.664818
2	CLOPE	0.7815	0.6872	0.5847	0.63182
3	**DBSCAN**	**0.9675**	**0.9132**	**0.8287**	**0.8689**
4	Filtered cluster	0.6754	0.6156	0.5843	0.599542
5	Farthest first	0.7532	0.7245	0.6789	0.700959
6	COWEB	0.8765	0.8567	0.7756	0.814135
7	K-Means clustering	0.9134	0.9072	0.7862	0.842377

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

8	CLARA	0.9078	0.8967	0.7432	0.812766

The obtained results show that DB SCAN performs well in terms all the metrics compared with other algorithms. The comparison among various clustering algorithms on the basis of time taken and number of clusters formed over the student dataset have been performed. For given dataset, EM algorithm took more time to perform clustering whereas farthest first algorithm took very less time. In case of clustered instances, DBSCAN algorithm formed larger amount of clusters whereas farthest first algorithm and filtered cluster algorithm formed less amount of clusters. So according to time taken, farthest first algorithm is preferred more than other algorithms and according to clustered instances, DBSCAN algorithm is preferred more than other algorithms.



The algorithm is also tested in terms of accuracy, precision, recall and f-measure. DBSCAN algorithm gives more accuracy and f-measure value when compared to all other algorithm. The chart above shows the performance of various clustering algorithm. The result shows that DBSCAN has higher performance in terms of all the metrics. Hence it can be conclude that DBSCAN outperforms all other algorithms for our student performance analysis dataset.

VII. CONCLUSION AND FUTURE WORK

The main aim of this paper is to make a detailed survey on different kind of clustering techniques its advantages, disadvantages and its applications. So this paper can be used as a quick review to know about the different clustering techniques available in data mining. This paper also compares the performance of different clustering algorithm using different metrics among which DBSCAN algorithm performs well in terms all the measure and so in future all our proposed algorithm will be based on improving this algorithm to produce improved results.

REFERENCES

1. R. S. J. D. Baker and K. Yacef, “The state of educational data mining in 2009: A review and future visions,” Journal of Educational Data Mining, vol. 1, no. 1, 2009.
2. J. Ranjan and K. Malik, “Effective educational process: A data mining approach, Vine, vol. 37, no. 4, pp. 502-515, 2007.
3. Lakshmanaprabu SK, K. Shankar, Ashish Khanna, Deepak Gupta, Joel J. P. C. Rodrigues, Plácido R. Pinheiro, Victor Hugo C. de Albuquerque, “Effective Features to Classify Big Data using Social Internet of Things”, IEEE Access, Volume.6, page(s):24196-24204, April 2018.
4. M. Miftakul Amin, Andino Maseleno, K. Shankar, Eswaran Perumal, R.M. Vidhyavathi, Lakshmanaprabu SK, “Active Database System Approach and Rule Based in the Development of Academic Information System”, International Journal of Engineering & Technology, Volume. 7, Issue-2.26, page(s): 95-101, June 2018.
5. Dwi AD Putra, Kamarul Azmi Jasmi, Bushrah Basiron, Miftachul Huda, Andino Maseleno, K. Shankar, Nur Aminudin, “Tactical Steps for E-Government Development”, International Journal of Pure and Applied Mathematics, Volume.119, No. 15, page(s): 2251-2258, June 2018.
6. Revathi S and Nalini T (2013) Performance Comparison of Various Clustering Algorithm. In International Journal of Advanced Research in Computer Science and Software Engineering, 3: 67-72
7. Tiwari M and Jha MB (2012) Enhancing the performance of Data Mining Algorithms in Letter Image Recognition Data. In International Journal of Computer Application in Engineering Sciences, 2: 217-220
8. Pallavi and Godara S (2011) A Comparative Performance Analysis of Clustering Algorithms. In International Journal of Engineering Research and Application, 1: 441-445
9. K. Ying, M. Chang, A. F. Chiarella, and J.-S. Heh, Clustering students based on their annotations of a digital text,” in Proc. 2012 IEEE Fourth Int. Conf. Technol. Educ., Jul. 2012, pp. 20-25.
10. Shankar, K., Lakshmanaprabu, S. K., Gupta, D., Maseleno, A., & de Albuquerque, V. H. C. (2018). Optimal feature-based multi-kernel SVM approach for thyroid disease classification. The Journal of Supercomputing, 1-16.
11. Eka Sugiyarti, Kamarul Azmi Jasmi, Bushrah Basiron, Miftachul Huda, K. Shankar, Andino Maseleno, “Decision Support System of Scholarship Grantee Selection using Data Mining”, International Journal of Pure and Applied Mathematics, Volume.119, No. 15, page(s): 2239-2249, June 2018.
12. P. Moreno-Clari, M. Arevalillo-Herraez, and V. Cerveron-Lleo, “Data analysis as a tool for optimizing learning management systems,” in Proc. Ninth IEEE Int. Conf. Adv. Learn. Technol., Jul.2009, pp. 242-246.
13. H. Grob, F. Bensberg, and F. Kaderali, “Controlling open source intermediaries-a web log mining approach,” IEEE Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews, vol. 1, pp. 233-242, 2004,
14. M. Pechenizkiy, T. Calders, E. Vasilyeva, and P. De Bra, “Mining the student assessment data: Lessons drawn from a small scale case study,” EDM, 2008.
15. Tri Susilowati, Kamarul Azmi Jasmi, Bushrah Basiron, Miftachul Huda, K. Shankar, Andino Maseleno, Anis Julia, Sucipto, “Determination of Scholarship Recipients using Simple Additive Weighting Method”, International Journal of Pure and Applied Mathematics, Volume.119, No. 15, page(s): 2231-2238, June 2018.
16. F. Getúlio, R. De Janeiro, F. G. V Online, M. A. Amaral, U.Universidade, P. Ffalm, and B. R. Km, “Analysing users, access logs in moodle to improve e learning analisando logs de acessos dos usuários do moodle para melhorar e-learning cássia blondet baroque alexandre barcellos joão carlos da silva freitas carlos juliano longo,”in Proc. 2007 Euro Am. Conf. Telemat. Inf. Syst., 2007, pp. 1-4.
17. C. Romero, S. Ventura, and E. García, “Data mining in course

management systems: Moodle case study and tutorial," Comput. Educ., vol. 51, no. 1, pp. 368-384, Aug. 2008.

18. B. M. I, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "T2a predicting student performance," 2003.
19. D. Ibrahim and Zaidah, Rusli, "Predicting students" academic performance: comparing artificial neural network, decision tree and linear regression," in Proc. the 21st Annual SAS Malaysia Forum, 2007, pp. 1-6.
20. T. Etchells, À. Nebot, A. Vellido, P. Lisboa, and F. Mugica, "Learning what is important: Feature selection and rule extraction in a virtual course," ESANN, pp. 26-28, 2006.
21. S. K. Lakshmanaprabu, K. Shankar, M. Ilayaraja, Abdul Wahid Nasir, V. Vijayakumar Naveen Chilamkurti, "Random forest for big data classification in the internet of things using optimal features", International Journal of Machine Learning and Cybernetics, January 2019. https://doi.org/10.1007/s13042-018-00916-z
22. Andino Maseleno, Alicia Y.C. Tang, Moamin A. Mahmoud, Marini Othman, Suntiaji Yudo Negoro, Soukaina Boukri, K. Shankar, Satria Abadi, Muhamad Muslihudin, "The Application of Decision Support System by Using Fuzzy Saw Method in Determining the Feasibility of Electrical Installations in Customer's House", International Journal of Pure and Applied Mathematics, Vol.119, No. 16, page(s): 4277-4286, July 2018.
23. A. S. Sabitha and D. Mehrotra, "User centric retrieval of learning objects in LMS," in Proc. 2012 Third Int. Conf. Computer. Commun. Technol., Nov. 2012, pp. 14-19.
24. K. Govindarajan, T. S. Somasundaram, and V. S. Kumar, "Particle swarm optimization (PSO)-based clustering for improving the quality of learning using cloud computing," in Proc. 2013 IEEE 13th Int.Conf. Adv. Learn. Technol., Jul. 2013, pp. 495-497.
25. N. Jyothi, K. Bhan, U. Mothukuri, S. Jain, and D. Jain, A recommender system assisting instructor in building learning path for personalized learning system," in Conf. on Technol. Educ. (T4E), 2012, pp. 228-230.
26. Muhammad Muslihudin, Risma Wanti, Hardono, Nurfaizal, K. Shankar, Ilayaraja M, Andino Maseleno, Fauzi, Dwi Rohmadi Mustofa, Muhammad Masrur, Siti Mukodimah, "Prediction of Layer Chicken Disease using Fuzzy Analytical Hierarcy Process", International Journal of Engineering & Technology, Volume. 7, Issue-2.26, page(s): 90- 94, June 2018.
27. Saranya, S., R. Ayyappan, and N. Kumar. "Student Progress Analysis and Educational Institutional Growth Prognosis Using Data Mining." International Journal of Engineering Sciences & Research Technology, 2014
28. Hicheur Cairns, Awatef, et al. "Towards Custom-Designed Professional Training Contents and Curriculums through Educational Process Mining." IMMM 2014, the Fourth International Conference on Advances in Information Mining and Management. 2014
29. Arora, Rakesh Kumar, and Dharmendra Badal. "Mining Association Rules to Improve Academic Performance." (2014).
30. Sukanya, M., S. Biruntha, Dr S. Karthik, and T. Kalaikumaran. "Data mining: Performance improvement in education sector using classification and clustering algorithm." In International conference on computing and control engineering, (ICCCE 2012), vol. 12. 2012.
31. Sakurai, Yoshitaka, Setsuo Tsuruta, and Rainer Knauf. "Success Chances Estimation of University Curricula Based on Educational History, Self-Estimated Intellectual Traits and Vocational Ambitions." Advanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on. IEEE, 2011.
32. Archer, Elizabeth, Yuraisha Bianca Chetty, and Paul Prinsloo. "Benchmarking the habits and behaviors of successful students: A case study of academic-business collaboration." The International Review of Research in Open and Distance Learning 15.1 (2014).
33. P. Arabie, L. J. Hubert, and G. De Soete. Clustering and Classification. World Scientific, 1996
34. Ji He, Man Lan, Chew-Lim Tan, Sam-Yuan Sung, Hwee-Boon Low, "Initialization of Cluster refinement algorithms: a review and comparative study", Proceeding of International Joint Conference on Neural Networks [C]. Budapest, 2004.
35. Biswas, G., Weingberg, J. and Fisher, D.H., ITERATE: A conceptual clustering algorithm for data mining. IEEE Transactions on Systems, Man, and Cybernetics. v28C. 219-230.
36. Uthayakumar, J., Metawa, N., Shankar, K., & Lakshmanaprabu, S. K. (2018). Financial crisis prediction model using ant colony optimization. International Journal of Information Management. https://doi.org/10.1016/j.ijinfomgt.2018.12.001
37. Lakshmanaprabu, S. K., Shankar, K., Gupta, D., Khanna, A., Rodrigues, J. J., Pinheiro, P. R., & de Albuquerque, V. H. C. (2018). Ranking analysis for online customer reviews of products using opinion mining with clustering. Complexity, 2018.
38. George Karypis, Eui-Hong (Sam) Han, Vipin Kumar, Chameleon: Hierarchical Clustering Using Dynamic Modeling, Computer, v.32 n.8, p.68-75, August 1999 [doi.10.1109/2.781637
39. Tian Zhang, Raghu Ramakrishnan, Miron Livny, BIRCH: an efficient data clustering method for very large databases, Proceedings of the 1996 ACM SIGMOD international conference on Management of data, p.103-114, June 04-06, 1996, Montreal, Quebec, Canada
40. https://en.wikipedia.org/wiki/Data_mining
41. Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". Published in Proceeding of 2nd international Conference on Knowledge Discovery and date Mining (KDD 96)
42. Uthayakumar, J., Metawa, N., Shankar, K., & Lakshmanaprabu, S. K. (2018). Intelligent hybrid model for financial crisis prediction using machine learning techniques. Information Systems and e-Business Management, 1-29. https://doi.org/10.1007/s10257-018-0388-9
43. Lydia, E. L., Kumar, P. K., Shankar, K., Lakshmanaprabu, S. K., Vidhyavathi, R. M., & Maseleno, A. (2018). Charismatic Document Clustering Through Novel K-Means on- negative Matrix Factorization (KNMF) Algorithm Using Key Phrase Extraction. International Journal of Parallel Programming, 1-19.
44. Yiling Yang, Xud ong Guan, Jinyuan You, "CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data", ISSN: 58113-567-X/02/0007 and Year 2002s
45. O. C. Santos and J. G. Boticario, "Modeling recommendations for the educational domain," Procedia Computer. Sci., vol. 1, no. 2, pp. 2793-2800, Jan. 2010.
46. Tiwari M and Singh R (2012) Comparative Investigation of K-Means and K-Medoid Algorithm of IRIS Data. In the International Journal of Engineering Research and Development, 4: 69-72
47. Nur Aminudin, Eni Sundari, K. Shankar, P. Deepalakshmi, Fauzi, Rita Irviani, Andino Maseleno, "Weighted Product and Its Application to Measure Employee Performance", International Journal of Engineering & Technology, Volume. 7, Issue-2.26, page(s): 102-108, June 2018.
48. Sharma N, Bajpai A and Litoriya R (2012) Comparison the various Clustering algorithms of Weka. In International Journal of Emerging Technology and Advanced Engineering, 2:73-80
49. S. Chen and X. Liu, "An integrated approach for modeling learning patterns of students in web-based instruction: A cognitive style perspective," ACM Trans. Computer. Interact. vol. 15, no. 1, 2008.
50. P. Golding and O. Donaldson, "Predicting academic performance," in Proc. Front. Educ. 36th Annual. Conf., 2006, pp. 21-26.
51. Andino Maseleno, Alicia Y.C. Tang, Moamin A. Mahmoud, Marini Othman, K. Shankar, "Big Data and E - Learning in Education", International Journal of Computer Science and Network Security, Vol.18 No.5, page(s): 171-174, May 2018.