

# An Empirical Study to Predict Students Performance related to Womens Educational Issues using Hierarchical and Fuzzy Clustering Algorithms

A. Sumathi, Ph.D Scholar, R&D Centre, Bharathiar University, Coimbatore,

Assistant Professor of Computer Science, Navarasam Arts & Science College for women, Erode,

India. sumiharsi@gmail.com

N. Sengottaiyan, Director, Sri Shanmugha College of Engineering and Technology, Salem, India,

nsriram3999@gmail.com

**Abstract** In the present world, one of the most important issues to be sorted out is the educational issues related to women. Due to many reasons, most of them discontinue their studies in the midway. Improper guidance and lack of knowledge with parents in the social related issues is one of the key problem. Many problems related to Educational Data Mining (EDM) for academic objectives have been reported in the recent literatures. Those problems includes the analysis of Student learning, cognitive learning, modelling, behaviour, risk and the individual performance. The main objective of this paper is to identify the social and personal issues for women related to the field of education and to provide the possible solutions. A survey on EDM has been carried out for better understanding. New data set creation, data analysis and implementation phase has been performed with the help of various clustering methods and the results are compared.

**Keywords** —Clustering, Data mining, Educational Issues, Hierarchical, K-Means, Social factors.

## I. INTRODUCTION

Educational Data Mining (EDM) is a rising field investing data in educational context by applying distinctive Data Mining (DM) techniques/tools. EDM acquires properties from areas like Learning Analytics, Statistics, Artificial Intelligence, Information Technology, Psychometrics, Machine learning, Database Management System, Data Mining and Computing. It can be considered as interdisciplinary research field which gives inborn information of teaching and learning process for successful education. The exponential development of educational data from heterogeneous sources results an urgent need for research in EDM. This can help to meet the objectives and to determine specific goals of education. EDM related to academic objective can be classified into Person oriented, Department/Institutions oriented and Domain oriented.

## II. RESEARCH BACKGROUND

The EDM is a repetitive and knowledge discovery process that holds hypothesis formulation, refinement and testing [1]. The process of converting the raw data into useful information that can create an impact on the practice of an educational research is referred as EDM [2]. For the past decades, many researches apply different algorithms to convert the data in to useful context related to education [3]. The authors Romero and Ventura in the year 2007 [4] carried

out a comprehensive research work that took between 1995 and 2005. This research work specifies the data mining application ranging from traditional educational institutions to web-based management system. The EDM works published between the year 2010 and 2013 were analyzed by Pena-Ayala[5]. All these works were focused on various algorithms, tasks related to education and methods. A non-parametric clustering technique to carry out the studies through online was presented by Zaiane & Luo [6]. This is specifically carried out for mining the offline web activity of the learners. Zaiane [7] studied the collaborative filtering method which is adopted for the effective e-learning development. Corbet & Wagner [8] implemented the method of prediction in a scientific manner in order to learn the surroundings by exploitation process rather than system learning methodology. Familiar tools that can be used for supporting EDM was provided by Brusilovsky & Peylo [9]. Those tools are used to study how the EDM prediction method are used for transforming the student models [10]. The student modeling is an outstanding research field in EDM [11]. Garcia et al [12] developed a toolkit that is helpful in operating the course management systems which is capable of providing the relevant information to the average users. Based on students' progress, Wang & Liao [13] proposed the DM techniques that are used to create dynamic learning exercises. Even though the educational institutions use large number of tools and analysis methods

for the purpose of accessing the course materials, the learners are not able to track the entire activity they perform [14]. Pham & Afify [15] presented the review article that addresses the various engineering applications that uses clustering algorithms to solve complex problems. Jain et al [16] discusses the pattern clustering methods which aims at providing the relevant advice and references to access fundamental concepts. Johannes & Andreas [17] discussed the overview of various techniques of clustering algorithms in data mining along with essential ingredients of the demographic cluster algorithm. Han [18] in the book Data Mining - Concepts and Techniques presented various methods used in the data mining along with the concepts and its related techniques. Jiyuan [19] et al presented the concept of visualization data mining for the distribution of high dimensional data. An article on different heuristics for many engineering problems has been presented by Narges & Dlanne [20]. Yifan Li et al [21] evaluated the cluster moving objects that captures the interesting pattern changes during the movement of clusters. Sherin et al [22] proposed the enhanced swarm-like agents in the data mining process which doesn't require initial partitioning seeds but it adopt the changes dynamically. Brun et al [23] presented the model-based system for the purpose of evaluation of clustering validation measures. By implementing this technique in the data mining process, it is found that the performance of the validity indices is highly variable. Vijayalakshmi et al [24] attempted to implement the data mining techniques for a social cause that dealt with the infertility in women and the analysis has been carried out with WEKA tool. Fahad et al [25] formulated the modified k means algorithm for big data clustering, which consumes very less time, more effective and efficient. Amjad Abu Saa [26] predicted the students performance based on the personal and social factors. Nelofar Rahman [27] presented the data mining techniques for business analysis to provide the effective solution in a quicker time. Jamuna et al [28] presented the SVM techniques to predict the students' performance and the comparison has been carried out with the existing algorithms. Anu Sharma et al [29] presented the survey on EDM techniques related to the field of social network analysis and it also dealt with the future trends in the research area of social related issues. Sagardeep Roy et al [30] epitomizes the review on data mining techniques to analyze the students' performance at various levels. This paper deals with the misaddressed issues related to womens education in this society. The analysis has been carried out with the R-tool with the data set collected from the students as the input.

### III. PROCESS OF DATA MINING

EDM is an emerging is connected with improving existing methods, explore new methods in identifying the data and implementing the identified new methods for better understanding of the performance of the students in various

aspects. The hypothesis formulation plays a major role in EDM, where it is used to create large volume of data. The crucial task in the EDM is the data creation and data validation process. This is also known as pre-processing of data. Once the pre-processing of data is completed, implementation of various tools on the processed data for the purpose of interpretation and the same will be provided to end users. Further recommendation will be suggested for the refinement of problems/task. In data mining, different sets of data is required for different process as it is task-oriented. Defining the given task and selecting the appropriate data from the available database is the primary task. Three different issues arises in the data selection. Setting up a clear and concise problem description is the first issue, identifying the relevant data is the second issue and the independence of the selected variables for the data selected is the third issue. Normally, types of data sources for business applications include demographic data (such as income, education, number of households, and age), socio-graphic data (such as hobby, club membership, and entertainment), transactional data (sales records, credit card spending, issued checks), and so on. The data type can be categorized as quantitative and qualitative data. Quantitative data is measurable using numerical values. It can be either discrete (such as integers) or continuous (such as real numbers). Qualitative data, also known as categorical data, contains both nominal and ordinal data.

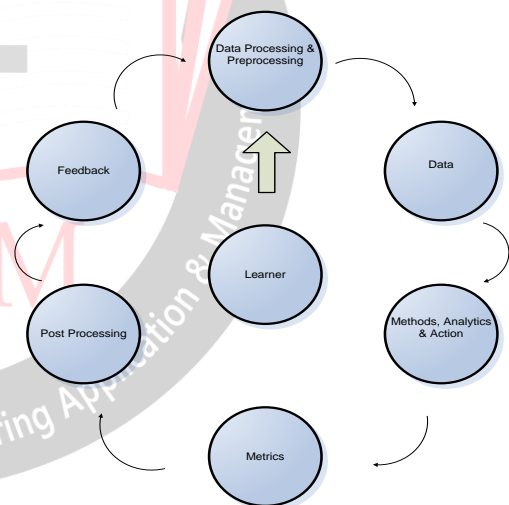


Figure 1. Process involved in Educational Data Mining

### IV. CLUSTERING METHODS

Clustering is referred as non-monitored learning technique for grouping similar data points. Clustering algorithm is the concept of assigning the large range of data points to smaller groups. The data points in the same cluster share similar properties whereas they are dissimilar in different groups. Image Segmentation, Grouping of web pages, Market Segmentation and Scientific Engineering Analysis are the few applications of Clustering [15]. The four major classifications of the clustering methods [16-23] are density-based methods, grid-base methods, hierarchical methods and partitioning methods. Apart from the

mentioned methods, few other techniques such as hierarchical, k-means and fuzzy clustering has been formed.

### A. K-Means Clustering

K-means clustering is one of the best clustering methods in the data mining process, where 'n' number of objects is being converted into 'k' clusters. The objects generally belongs to the cluster with the nearest mean. The output of this method will be the defined k different clusters with inherent characteristics. Based on the similarity, the data points are clustered together. The inputs for this algorithm are the data sets and the number of clusters 'K'. Collection of each data points feature is referred as data set and upon repetitive procedure, the end results are obtained. For a value of 'K' centroids, the initialization of the algorithm takes place, where the value of k can be generated in a random structure or from the data set. Let the centroid collection in Set C is  $C_i$ , then the data point 'x' is assigned to cluster based on the equation

$$c_i = \underset{c_i \in C_1}{\operatorname{argmin}} \operatorname{dist}(c_i, x)^2 \quad (1)$$

where  $\operatorname{dist}(\cdot)$  represents the standard Euclidian distance. Let  $S_i$  represent the set of data point assignments for each  $i^{\text{th}}$  cluster. The centroid points are recomputed in the centroid update step. This is carried out by calculating the mean points assigned to that particular cluster. This is represented by the equation

$$C_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i \quad (2)$$

The iteration process is carried out by the algorithm between the aforementioned steps until it reaches the stopping criteria. On execution, convergent to a result is guaranteed, in which it may be the local optimum. Executing the algorithm for many iterations with random structures might give a possible and expected result. Choosing 'K' is one of the important aspect in the K-means clustering. To determine the count of clusters, the end user must run the algorithm for different range of K values and the obtained results are to be compared. The results of different ranges of K values can be compared by calculating the mean distance between the data points and the cluster centroid. As this method decreases the metric, it is not opted for determining the solution. To overcome the aforesaid problem, a function 'K' is plotted based on the mean distance to the centroid and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine K. Many other techniques are available for determining the value 'K'. Cross-validation, Jump method, information criteria, G-means algorithm and silhouette method falls under this category.

### B. Hierarchical Clustering

The process of building the hierarchy of clusters is referred as Hierarchical clustering. Here the data points are assigned to each cluster of its own. Merging of two nearest cluster takes place. The algorithm gets terminated when only single cluster is left out. Nearest clusters are paired together and this process continues till one cluster remains at the top end. The distance between two clusters in the data space is determined by the height in the dendrogram where two clusters are merged. The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram.

Two important things to be studied in hierarchical clustering are as follows.

- This algorithm can be implemented in top-down and bottom-up approach only when all the data points are assigned in the same cluster and the performance is carried out till all the data points are assigned to single cluster.
- Based on the closeness of the cluster, the decision on merging of clusters takes place.

### C. Fuzzy Clustering

This is one of the mostly used algorithm in the process of EDM. The major difference in FCM when compared to other algorithms is that the membership of a particular data point is not decided by itself, rather it calculates the degree of membership of a data point belonging to a particular cluster. This algorithm is developed by Dunn and improved by Bezdek. Here predetermination of clusters play a major role. Therefore it can be said that the accuracy of the clustering is required and so the required tolerance measures can be placed. As the calculation of membership is viable, this method is faster than the other algorithms. The required accuracy is obtained by the repeated iterations with a specific clustering exercise. The fuzzy clustering aims at minimizing the objective function given by the equation 3

$$J = \sum_{i=1}^N \sum_{j=1}^c \delta_{ij} \|x_i - C_j\|^2 \quad (3)$$

Here, N - number of data points

C - number of clusters required

$C_j$  - centre vector for cluster j

$\delta_{ij}$  is the degree of membership for the  $i^{\text{th}}$  data point  $x_i$  in cluster j.

$\delta_{ij}$  represents the degree of membership and it is given by the equation

$$\text{Error! Reference source not found.} \quad (4)$$

where m-fuzziness coefficient

The centre vector  $C_j$  is calculated by the equation

$$\text{Error! Reference source not found.} \quad (5)$$

The FCM algorithm determines the expected accuracy of the degree of the membership. When the iteration moves from one to another, the accuracy is calculated considering the largest value of the entire cluster. Assume 'k' be the initial iteration and 'k+1' be the next iteration. Then the measure of accuracy between the iterations is calculated as follows.

$$\text{Error! Reference source not found.} \quad (6)$$

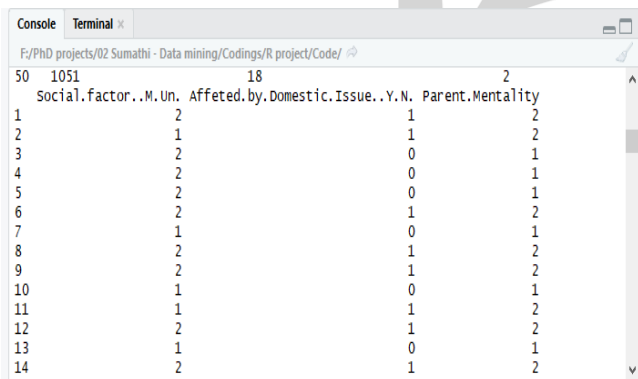
$\delta_{ij}^{(k+1)}$  and  $\delta_{ij}^k$  represents the degree of membership at iteration k and k=1 respectively.

## V. IMPLEMENTATION USING R-TOOL

R tool is an open source tool specifically used for statistical computing and graphical purpose. This tool is mostly prescribed by the data miners for the analytical purpose. The usage of R tool has increased to a higher level. As per the survey taken in June 2018, the R tool holds 10th place in the TIOBE index. It is a GNU package and the source code is written in C, Fortran and R. While R has a command line interface, there are several graphical front-ends, most notably RStudio and RStudio Server, which are the only GUIs developed by the R Foundation. The real time dataset has been created and analyzed. The data has been collected



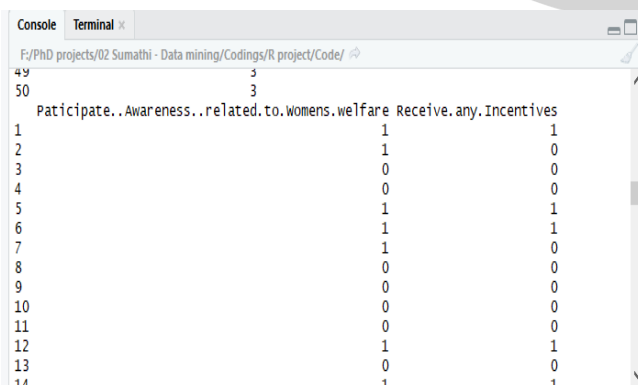
from the under graduate students of computer science and information technology at Navarasam Arts and Science College for Women, Erode district, Tamilnadu. A total of 100 samples of parameters has been collected. The implementation has been carried out with the collected samples using Data Mining R package. In R tool, all the variables, data and functions are stored in the type of named objects. Various social related issues has been taken as parameters for the purpose of analysis. Fuzzy clustering and hierarchical clustering analysis has been carried out for the data set created. The following social and personal related issues are opted for the purpose of analysis. Age when Join the Course, Course Related to High school Major, Social factor (M/Un), Affected by Domestic Issue (Y/N), Parent Mentality, Educational facilities (G/B/W), Participate Awareness related to womens welfare, Receive any Incentives, Worked as a Labour, Facing Financial Difficulty, Physically Challenged (or) Any Health Problem, Attendance, and End semester Result. The attributes selected for the purpose of analysis plays a major role in the field of womens education. The corresponding attributes and its possible values are assigned the possible numeric values. Fuzzy and hierarchical clustering analysis has been presented in this article.



	Social.factor..M.Un.	Affeted.by.Domestic.Issue..Y.N.	Parent.Mentality
1	2	1	2
2	1	1	2
3	2	0	1
4	2	0	1
5	2	0	1
6	2	1	2
7	1	0	1
8	2	1	2
9	2	1	2
10	1	0	1
11	1	1	2
12	2	1	2
13	1	0	1
14	2	1	2

Figure 2. Data set with values assigned for domestic issue in R-tool

Figure 2 depicts the values assigned to the attributes such as social factor, affected the domestic issue and parent mentality. The social factor carries the possibility of getting married or unmarried, the domestic issue holds the possible value of either Yes or No and the parent mentality gets the value of supportive or discourge.



	Participate..Awareness..related.to.womens.welfare	Receive.any.Incentives
1	1	1
2	1	0
3	0	0
4	0	0
5	1	1
6	1	1
7	1	0
8	0	0
9	0	0
10	0	0
11	0	0
12	1	1
13	0	0
14	1	1

Figure 3. Data set with values assigned to receive any incentive issue in R-tool

Figure 3 depicts the values assigned to the attributes such as participating in awareness programmes related to womens

welfare and receive any incentives which hold the possible value Yes or No.

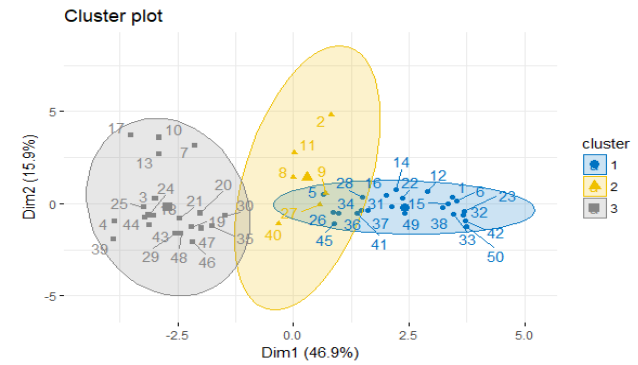


Figure 4. Output obtained from fuzzy clustering with three different cluster group

Figure 4 illustrates the cluster result obtained for three different clusters formed with similar set of values. From the figure it is clearly understood that the mid-cluster carries very less number of data points as it depends upon the input data values assigned to particular social factor.

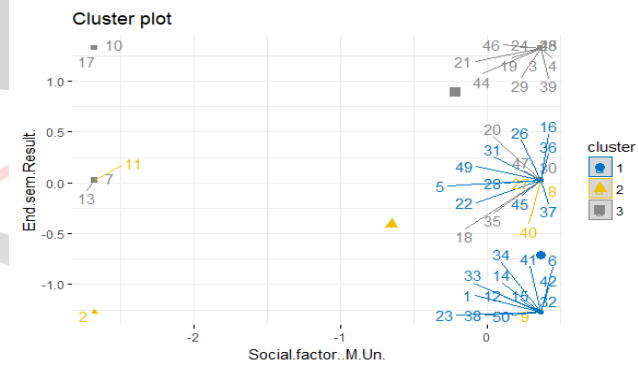


Figure 5. Output obtained from fuzzy clustering for social factor vs end semester result

The performance analysis between the social factor versus the end semester result is shown in figure 5. It is observed that the social factor carries the value of married or unmarried plays a major role in the discontinuity of the course in their mid-way. It is expressed with the cluster formation group 2 and 3 (shown in figure 5). The prediction of the social factors that affect the women education is discussed in this section.

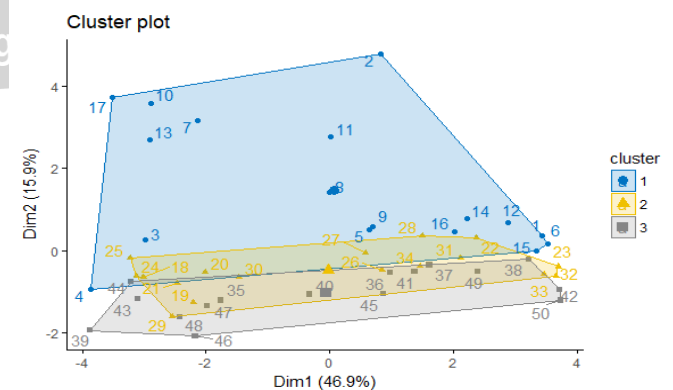


Figure 6. Hierarchical Cluster representation for 50 samples with 3 clusters

Figure 6 & 7 illustrates the hierarchical cluster representation and its analysis on the factor map with three different clusters. From both the figures, it is seen that the

cluster 2 exhibits very lesser data count when compared to other groups.

Hierarchical clustering on the factor map

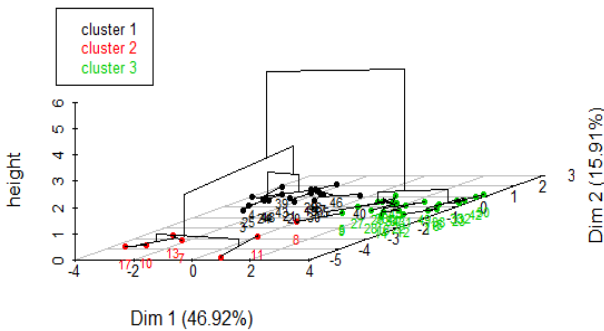


Figure 7. Hierarchical Cluster representation on the factor map with three different clusters

The representation of 100 and 50 samples selected for the purpose of analysis has been pictured from figure 4 to figure 7. Figure 4 and 5 represents the fuzzy clustering based analysis and figure 6 & 7 represents the hierarchical clustering based analysis. It is observed that from the attributes chosen, the domestic issue affects the womens education to a greater extent when compared with other attributes.

Table 1 Percentage of accuracy related to personal and social factors

Factors	PA	HC	FC	Probability
Personal Factors	92 %	91 %	92 %	0.0711
Social Factors	98 %	98.2 %	98.7%	0.1996

PA - Prediction Accuracy; HC - Hierarchical Clustering; FC - Fuzzy Clustering

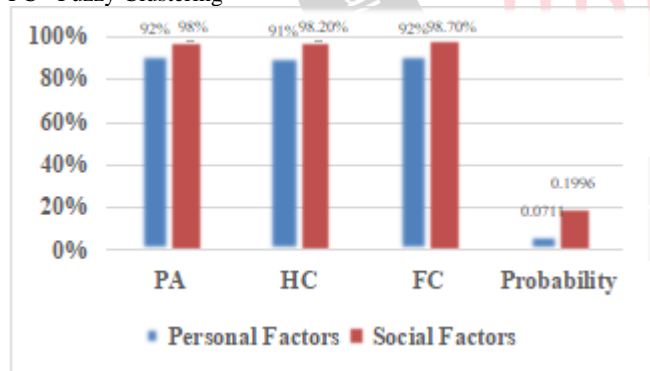


Table 1 epitomizes the accuracy level obtained with the help of hierarchical and fuzzy clustering methods for the personal and social factors taken as the main cause for the retardation of study in case of women children. It is seen that the fuzzy clustering being a rule base system, produces 92% of accuracy level where as in hierarchical clustering it is only 91%. When social factor is concerned, the accuracy level is 98.7% with fuzzy clustering and it is only 98.2% with hierarchical clustering. The social factors have the highest priority over the personal factors, also which is inferred from table 1.

## VI. CONCLUSION

In this paper, real time data set has been framed, analyzed, categorized and implemented with the help of two effective algorithms under clustering method. 12 attributes has been chosen for analyzing the retardation of womens education. Among the attributes chosen, the effect of physically handicapped shows only a minimal effect as the count of that attribute is only 5% of the total sample collected. Many of the other attributes are also found to be more effective on working. Other attributes has degree of effect on performance prediction. Performance of the students at the end of the semester is also analyzed with the different data mining methods. The fuzzy clustering method provides 92% of accuracy level pertaining to personal factors and it is of 98.7% with social factors. The prediction accuracy level is 92% with personal factors and it is 98% related to social factors. Probability of the values plays a major role in the prediction. From table 1 it is seen that the probability of the social factor is 0.1996 which is higher than the personal factor. From the above prediction it is inferred that the social factor plays a major role in stopping the women children from continuing their education. Using this prediction, the improvement in the educational system pertaining to womens education can be improved and at the same time the quality of the education in all aspects can be improved.

## ACKNOWLEDGMENT

I thank the management, staff and students of Navarasam College of Arts and Science for Women, Erode for their continuous support rendered towards this research work.

## REFERENCES

- [1] A. A. Al-shargabi and A.N. Nusari, "Discovering Vital Patterns From UST Students Data by Applying Data Mining Techniques," IEEE Int. Conf. Computer and Automation Engineering, China, 2010, pp. 547-551.
- [2] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," Systems, Man, and Cybernetics, Part C: IEEE Transactions on Applications and Reviews, 2010, pp. 601-618.
- [3] M. F. M. Mohsin, N. M. Norwawi, C. F. Hibadullah, and M. H. A. Wahab, "Mining the student programming performance using rough set," presented at the 2010 Int. Conf. Intelligent Systems and Knowledge Engineering.
- [4] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Systems with Applications, 2007, vol. 33, pp. 135-146.
- [5] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," Expert Systems with Applications, 2014, vol. 41, pp. 1432-1462.

- [6] O. R. Zaiane and J. Luo, "Web usage mining for a better web-based learning environment," in Proc. Conf. advanced technology for education, 2001, pp. 60-64.
- [7] O. R. Zaiane, "Building a recommender agent for e-learning systems," *Int. Conf. Computers in Education*, 2002, pp. 55-59.
- [8] R. S. Baker, A. T. Corbett, and A. Z. Wagner, "Off-task behavior in the cognitive tutor classroom: when students game the system," presented at *Proc. of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 383-390.
- [9] P. Brusilovsky and C. Peylo, "Adaptive and intelligent web-based educational systems," *International Journal of Artificial Intelligence in Education*, vol. 13, pp. 159-172, 2003.
- [10] J. E. Beck and B. P. Woolf, "High-level student modeling with machine learning," *Intelligent tutoring systems*, 2000, pp. 584-593.
- [11] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, 2009.
- [12] E. García, C. Romero, S. Ventura, and C. de Castro, "A collaborative educational association rule mining tool," *The Internet and Higher Education*, 2011, vol. 14, pp. 77-88.
- [13] Y. H. Wang and H. C. Liao, "Data mining for adaptive learning in a TESL-based e-learning system," *Expert Systems with Applications*, vol. 38, pp. 6480-6485.
- [14] M. E. Zorrilla, E. Menasalvas, D. Marin, E. Mora, and J. Segovia, "Web usage mining project for improving web-based learning sites" *Computer Aided Systems Theory-EUROCAST*, 2005, pp. 205-210.
- [15] D. T. Pham, and A. A. Afify, "Clustering techniques and their applications in engineering". *Proc. Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 2010.
- [16] A.K. Jain, M.N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, 1999, vol.31.
- [17] J. Grabmeier and A. Rudolph, "Techniques of cluster algorithms in data mining," *Data Mining and Knowledge Discovery*, 2002, vol. 6, pp. 303-360.
- [18] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Academic Press, San Diego, California, USA, 2001.
- [19] Jiyuan An, Jeffrey Xu Yu, Chotirat Ann Ratanamahatana and Yi-Ping Phoebe Chen, "A dimensionality reduction algorithm and its application for interactive visualization," *Journal of Visual Languages and Computing*, 2007, vol.18, pp.48-70.
- [20] Nargess Memarsadeghi, Dianne P. O'Leary, "Classified Information: The Data Clustering Problem," *Computing in Science and Engineering*, 2003, vol.5, pp.54-60.
- [21] Yifan Li, Jiawei Han, Jiong Yang, "Clustering moving objects," *Proc. the tenth ACM SIGKDD International conf. Knowledge discovery and data mining*, 2004, Seattle, WA, USA. pp. 22-25,
- [22] M. Sherin, Youssef, Mohamed Rizk, Mohamed El-Sherif, "Enhanced swarm-like agents for dynamically adaptive data clustering," *Proc. of the 2nd WSEAS Int. Conf. Computer Engineering and Applications*, January 25-27, Acapulco, Mexico, *International Journal of Database Management Systems*, vol.3, 2011, pp.213-219.
- [23] Marcel Brun, Chao Sima, Jianping Hua, James Lowey, Brent Carroll, Edward Suh, Edward R. Dougherty, "Model-based evaluation of clustering validation measures," *Pattern Recognition*, 2007, vol.40, pp. 807-824.
- [24] N. Vijayalakshmi and UmaMaheswari, "Data Mining to elicit predominant factors causing infertility in Women", *International Journal of Computer Science and Mobile Computing*, vol. 5, 2016, pp. 5-9.
- [25] Ahammad Fahad and Mahbub Aslam, "A Modified K-Means Algorithm for Big Data Clustering", *International Journal of Computer Science & Engineering Technology*, vol. 6, 2016, pp. 129-132.
- [26] Amjad Abu Saa, "Educational Data Mining & Students' Performance Prediction", *International Journal of Advanced Computer Science and Applications*, vol. 7, 2016, pp. 212-220.
- [27] Nelofar Rahman, "Data Mining Techniques Methods Algorithms and Tools", *International Journal of Computer Science and Mobile Computing*, vol.6 Issue.7, 2017, pp. 227-231.
- [28] M. Jamuna and S. A. Shoba, "Educational data mining & students performance prediction using SVM techniques", *International Research Journal of Engineering and Technology*, vol. 4, 2017, pp. 1248-1254.
- [29] Anu Sharma, M.K. Sharma and R. K. Dwivedi, "Literature Review and Challenges of Data Mining Techniques for Social Network Analysis", *Advances in Computational Sciences and Technology*, vol. 10, 2017, pp. 1337-1354.
- [30] Sagardeep Roy and Anchal Garg, "Analyzing Performance of Students by Using Data Mining Techniques-A literature Survey", 4<sup>th</sup> International Conference on Electrical, Computer and Electronics, IEEE Uttarpradesh Section, 2017, pp. 130-133.