Journal of Computer Science 7 (11): 1652-1658, 2011
ISSN 1549-3636
© 2011 Science Publications

Fuzzy Modeled K-Cluster Quality Mining of Hidden Knowledge for Decision Support

[1]S. Prakash Kumar and [2]K.S. Ramaswami
[1]Department of Computer Applications,
Erode Sengunthar Engineering College, Erode-638 057, Tamilnadu, India
[2]Department of Mathematics, Coimbatore Institute of Technology,
Civil Aerodrome Post Coimbatore-641 014, Tamilnadu, India

Abstract: Problem statement: The work presented Fuzzy Modeled K-means Cluster Quality Mining of hidden knowledge for Decision Support. Based on the number of clusters, number of objects in each cluster and its cohesiveness, precision and recall values, the cluster quality metrics is measured. The fuzzy k-means is adapted approach by using heuristic method which iterates the cluster to form an efficient valid cluster. With the obtained data clusters, quality assessment is made by predictive mining using decision tree model. Validation criteria focus on the quality metrics of the institution features for cluster formation and handle efficiently the arbitrary shaped clusters. **Approach:** The proposed work presented a fuzzy k-means cluster algorithm in the formation of student, faculty and infrastructural clusters based on the performance, skill set and facilitation availability respectively. The knowledge hidden among the educational data set is extracted through Fuzzy k-means cluster an unsupervised learning depends on certain initiation values to define the subgroups present in the data set. **Results:** Based on the features of the dataset and input parameters cluster formation vary, which motivates the clarification of cluster validity. The results of quality indexed fuzzy k-means shows better cluster validation compared to that of traditional k-family algorithm. **Conclusion:** The experimental results of cluster validation scheme confirm the reliability of validity index showing that it performs better than other k-family clusters.

Key words: Decision support, fuzzy k-cluster, quality mining

INTRODUCTION

Nowadays many educational systems generate mountains of administrative data about students, courses, staff including lecturers, infrastructure, managerial systems. This data is a strategic resource for educational institution. Making the most use of these strategic resources will lead to the main objective of educational system, which is improving the quality of processes. To retain qualified in educational domain, a deep understanding of the knowledge hidden among the data is required. In today's education lack of deep and enough knowledge among the processes such as evaluation, counseling prevents management system from achieving this quality objective, so there has not been an efficient and effective use of their strategic resources yet. It results in extracting greater value from the raw data set and making use of strategic resources efficiently and effectively. It finally improves the quality of educational processes.

The ability of data mining improving the quality of educational processes by offering an enhanced version of newly proposed analysis model (DM_EDU) presented in (Delavari *et al*., 2005) used for the application of data mining in educational system. In addition to that, our main contribution is considering the quality improvement of decision-making processes. Specifically, the adopted approach uses an existing Multi Criteria Decision Making (MCDM) method, giving emphasis on the formulation of the process (Karakosta *et al*., 2008) as to be relatively straightforward to incorporate direct stakeholders' preferences. This article aims to develop a fuzzy Multicriteria Decision Making (MCDM) tool that equips with Analytic Hierarchy Process (AHP) framework (Cheong *et al*., 2008) to help users in semi-structured and unstructured decision making tasks. The methodology is based on CRISP-DM methodology Cross Industry Standard Process for Data Mining. In

Corresponding Author: S. Prakash Kumar, Department of Computer Applications, Erode Sengunthar Engineering College, Erode-638 057. Tamilnadu, India

practice, one may classify each item in more than two categories such as “bad”, “medium”, “good” and “excellent” (Amirzadeh *et al*., 2008). Based on this, we introduce a Fuzzy Multinomial chart (FM-chart) for monitoring a multinomial process.

Institutional data mining: Data mining is the process of autonomously extracting useful information or knowledge from large data stores or sets. Data mining consists of more than collecting and managing data, it also includes analysis and prediction. These tools can include statistical models, mathematical algorithms and machine learning methods such as neural networks or decision trees. Data mining is popularly known as knowledge discovery databases. As the educational systems are capable of collecting large amount of students profile data, data mining and rough set techniques can be applied to find interesting relationships between attributes of students.

The concept of data mining, involves three steps i.e., capturing and storing the data, converting the raw data into information and converting the information into knowledge. Data in this context comprises all the raw material an institution collects via normal operation. Capturing and storing the data is the first phase that is the process of applying mathematical and statistical formulas to “mine” the data warehouse. Mining the collected raw data from the entire institution may provide new information as to how students, parent’s and the institutions own processes really perform. Converting the raw data into information is the second step of data mining. Our survey on the current works in data mining field shows that one of the application domains that can take advantage of data mining benefits in education. Student Information System data is involved with three kinds of large data sets:

- Educational resources such as student databases, fees collection and individualized problems designed for use on assignments and examinations
- Information about users who create, modify, assess, or use these resources
- Activity log databases which log actions taken by students in behavioral characteristics and exam results

Educational institution quality assessment model: An item soon to be integrated in many educational systems is adoption of data mining. It can be best explained as the process of extracting useful knowledge and information including, patterns, associations, changes, anomalies and significant structures from a great deal of data stored in databases, data warehouses, or other information repositories. The genetic algorithm was applied to calibrate the fuzzy set model (Thongwan *et al*., 2011). Prior to the great usages that this technology brings into many application areas such as biomedical and DNA analysis, retail industry and marketing, telecommunications, web mining and recently has also been an interesting area of research in educational domain. This information overload also exists in the biomedical field, (Feldman *et al*., 2003) where scientific publications and other forms of text-based data are produced at an unprecedented rate.

Cluster validation on quality metrics: Requirements for the evaluation of clustering result, is well known in the research community and a number of efforts have been made especially in the area of pattern recognition. The most frequently used clustering is the K-family clusters, among which a variant of K-means cluster presented by evaluated a strategy of partial distance logic to k-means algorithm which avoids unnecessary distance calculation made by traditional k-means clusters. Their work was not complementary for different type of distribution. They initiated two ways for generating input data points i.e., normal and uniform distribution. The Author also used clustering for aggregating data in multiple tables for handling classification in relational databases which comprises of user control clustering and automatic non-overlapping clustering in multiple instances with genetic algorithm. Graph theory has been used in protein sequence (Jaber *et al*., 2009) clustering as a means of partitioning the data into groups, where each group constitutes a cluster.

They employed time-invariant and time-variant fuzzy time-series which cluster the data, find each cluster membership values, define and partition universe of discourse, then fuzzily historical data and logical relation and finally calculate forecasted outputs which increase the performance of cluster evaluation. This motivates us to adapt fuzzy clustering for predictive decision support on evaluating institutional data.

However, the issue of cluster validity is rather under-addressed in the area of databases and data mining applications, even though recognized as important. In general terms, there are approaches to investigate cluster validity (Xie and Beni, 1991). A cluster validity index for crisp clustering attempts to identify compact and well-separated clusters. The implementation of most of these indices is very computationally expensive, especially when the number of clusters and number of objects in the data set grows very large. With cluster validity on institutional quality assessment, we presented an efficient cluster quality validation index to measure the performance of institutional quality in terms of metrics such as performance students, capabilities and skill of the trainers and the infrastructure requirements.

MATERIALS AND METHODS

The institutional quality assessment model presented a cluster evaluation process on the metrics of student performance, faculty skill sets and infrastructural requirement. Then present a predictive process using decision tree model to evaluate the overall performance of educational system. In first process of cluster formation, provide a full access to the institution's educational records. With this access, they are able to evaluate the problems presented in the course after the students have used the educational materials, through some statistical reports. It also provides a quick review of students' submissions for every problem in a course. The instructor may monitor the number of submissions of every student in any assignment set and its problems. The total numbers of solved problems in an assignment set as compared with the total number of solved problems in a course are represented for every individual student. The proposed model in our work is shown in the Fig. 1 indicating the process adapted to evaluate decision support for institutional quality assessment.

The important task of the feedback tools for the instructor is to help identify the source of difficulties and the misconceptions students have about a topic. There are basically three ways to look at such homework data: by student, by problem, or crosscutting (per student, per problem). The amount of data gathered from large enrollment courses with over 200 randomizing assignment problems, each of them allowing multiple attempts, can be overwhelming.

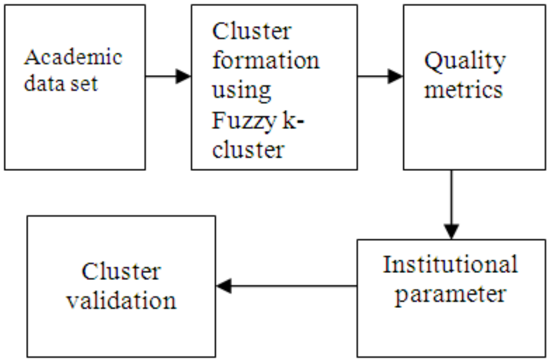

Fig. 1: Process framework for fuzzy modeled K-cluster quality mining

Every part of a multi-part problem is distinguished as a separate problem. The multi-instance problem is also considered separately, because a particular problem or one part of it might be used in different assignment sets. Finally, a Table 1 is created which includes all computed information from all students, sorted according to the problem order. In this step, the system has provided the following statistical information:

- Number of students: Total number of students who take a look at the problem. (Let number of students is equal to n)
- Tries: Total number of submissions to solve the problem:

$$\sum_{i=1}^{n} x_i$$

where, Xi denote a student try.

- Mod: Mode, maximum number of submissions for solving the problem
- Mean: Average number of the submissions:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- YES: Number of students solved the problem correctly
- Yes: Number of students solved the problem by override. Sometimes, a student gets a correct answer after talking with the instructor. This type of correct answer is called "corrected by override"
- %Wrng: Percentage of students tried to solve the problem but still incorrect:

$$100 * \left(\frac{n-(YES+yes)}{n}\right)$$

- SD: Standard Deviation of the students' submissions:

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Table 1: Cluster instances

Number of Clusters	Existing	Proposed
0	326	402
1	342	427
2	356	478
3	373	549
4	386	596
5	398	674

Fuzzy K-means cluster: Fuzzy K-Means is an extension of K-Means, the popular simple clustering technique. While K-Means discovers hard clusters (a point belong to only one cluster), Fuzzy K-Means is a more statistically formalized method and discovers soft clusters where a particular point can belong to more than one cluster with certain probability. In this study, we propose a new strategy to accelerate the k-means clustering algorithm (Al-Zoubi *et al*., 2008) through the Partial Distance (PD) logic.

The Fuzzy K-means accepts an input file containing vector points student and faculty data sets. The quality assessment model provides the cluster centers as input and /or allow canopy algorithm to run and create initial clusters. The proposed algorithm doesn't modify the input directories. Fuzzy K-Means Mapper reads the input cluster during its configure method, then computes cluster membership probability of a point to each cluster. Cluster membership is inversely proportional to the distance. Distance is computed using student assignment submission date, grade as distance measure. Output key is encoded cluster. Output values are the probability Value, input Point. Fuzzy K-Means Combiner receives all key: value pairs from the mapper and produces partial sums of the cluster membership probability times input vectors for each cluster. Output key is the encoded cluster. Output value is sum of cluster membership values in the partial sum, partial sum vector summing all such points. Fuzzy K-Means Reducer receives certain keys and all values associated with those keys. The reducer sums the values to produce a new centroid for the cluster which is output. Output key is the encoded cluster identifiers.

Cluster validity: The clustering validity criteria are classified into internal, external and relative. The proposed work focus on the relative association of faculty, students and infrastructure relative criteria is used as the validity measure. The process of cluster validation defines a relative validity index, for assessing the quality of partitioning for each set of the input values. The proposal formalize clustering validity index based on clusters' compactness (in terms of cluster density) and clusters' separation (combining the distance between clusters and the inter-cluster density). Cluster Density (ID) evaluates the average density in the region among clusters. The goal is the density in the area among clusters to be significant low. Then, considering a partitioning of the data set into more than two clusters (i.e., c>1) the inter-cluster density is defined as follows Eq. 1:

$$\mathrm{Inter\_dens}(c)=\sum_{i=1}^{c}\sum_{\substack{j=1\\ i\neq j}}^{c}\left(\frac{\begin{array}{c}d(\mathrm{clos\_rep}_i\\ ,\mathrm{clos\_rep}_j)\end{array}}{\mathrm{stdev}_i+\mathrm{stdev}_j},\mathrm{density}(u_{ij})\right),\ c>1, c\neq n \qquad (1)$$

where, clos_rep$_i$, clos_rep$_j$ are the closest representative points between clusters i and j and n the number of points in a data set. Also, u_{ij} is the middle point of the line segment defined by the closest clusters' representatives clos_rep$_i$, clos_rep$_j$. The term density (u_{ij}) is defined as Eq. 2:

$$\mathrm{density}(u_{ij})=\frac{\sum_{i=1}^{n_i n_j} f(x_i,u_{ij})}{n_i+n_j} \qquad (2)$$

where, clos_rep$_i$, clos_rep$_j$ are the closest representative points between cluster c_i and c_j and n the number of points in a data set. It represents the percentage of points in the cluster i and the cluster j that belong to the neighborhood of u_{ij}. The neighborhood of a data point, u_{ij}, is defined to be a hyper-sphere with center u_{ij} and radius the average standard deviation of the clusters between which we estimate the density. Also, the function f(x, u_{ij}) is defined as Eq. 3:

$$f(x,u_{ij})=\begin{cases}0, & \text{if } d(x,u_{ij})>(\mathrm{stdev}_i+\mathrm{stdev}_j)/2\\ 1, & \text{otherwise}\end{cases} \qquad (3)$$

It is obvious that a point belongs in the neighborhood of u_{ij} if its distance from u_{ij} is smaller than the average standard deviation of clusters. However, the actual area between clusters, whose density we are interested to estimate, is defined to be the area between the closest representative points. Clusters' Separation (CS) evaluates the separation of clusters taking into account both the distances between the closest clusters and the Inter-cluster density. The goal is the distances among clusters to be high while the density in the area among them to be low. Then, the clusters' separation is given by Eq. 4:

$$\mathrm{Sep}(c)=\frac{\sum_{i=1}^{c}\sum_{\substack{j=1\\ i\neq j}}^{c}\min\{d(\mathrm{clos\_rep}_i,\mathrm{clos\_rep}_j)\}}{1+\mathrm{Inter\_dens}(c)},\ c>1 \qquad (4)$$

where, clos_rep$_i$, clos_rep$_j$ are the closest representative points between clusters c_i and c_j.

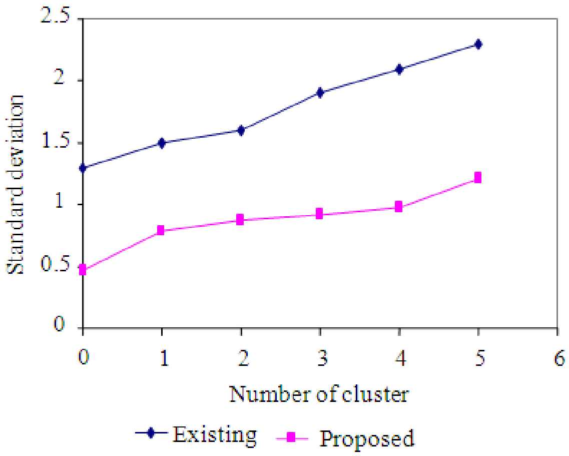

Fig. 2: Standard deviation with number of clusters verifying the Fuzzy K-Cluster Validity

Institutional parameters: The parameter used in the quality assessment of the educational system are listed down along with its data collections sources.

Types of informational measures:

- Student profile
- Faculty profile
- Curriculum
- Outcome profile
- Learning path ways
- Infrastructural facilities

Analytical evaluation:

- Analysis of student
- involvement and engagement
- Staff Performance relating
- to student results and career
- Resource Facility relation
- to easy and effective
- knowledge acquisition
- Quality assessment in terms of faculty performance, student outcome and resource availability
- Organizational change

The institution offers programs that take into consideration the social, cultural, economic and developmental needs of the country at local, regional and national levels, as well as the need for the country to compete effectively in global markets. The institution is valued as a partner by other higher education institutions, professional, government and non-government organizations; and industry, within the India and internationally. The institution is valued by its local community as provider of extension programs that are responsive to the needs of the community for people empowerment and self-reliance. On data integration, the information about the students, lecturers and courses stored in various tables are merged to have different information about lecturer and student course performance, academic and personal information of lecturers gathered together for each single student object.

RESULTS

The experimentation conducted on the student assessment and the faculty performance based on the results of the students and their profile two-step K-means fuzzy cluster technique are evaluated for its tolerance of diverse data types and user-friendly groupings. To establish typologies, in which case, far more manual categorization should have occurred prior to actual modeling. One way of understanding groupings typically involves examining a secondary level of factors associated with the main outcomes of the data mining project. This would mean going beyond persisting and non-persisting, transfer and non-transferred to a level that define when or how the outcome happened, for example, number of terms prior to a student became transfer ready, or number of courses continually taken by a student prior to becoming transfer ready. The following resultant clustering analysis represents a general analysis of the entire population to seek major centroids of student performance and staff performance. Since data mining is iterative work, this part of the analysis may occur before predictive modeling is conducted, so that somewhat homogenous populations exist to make the predicted score more precise. Fig. 2 shows that the standard deviation is efficient in proposed system when compared to the proposed system

The complexity of the validity index CDbw, is based on the complexity of its two terms cluster density and separation. Assuming d the number of attributes (data set dimension); c is the number of clusters, n is the number of database tuples; r the number of a cluster's representatives. Then the complexity of selecting the closest representative points of c clusters is $O(dc^2r^2)$. The intra-cluster density complexity is O(ncrd) while the complexity of inter-cluster density is $O(ndc^2)$. Then CDbw complexity is $O(ndr^2c^2)$. Usually, c, d, r<< n, therefore the complexity of our index for a specific clustering scheme is O(n).

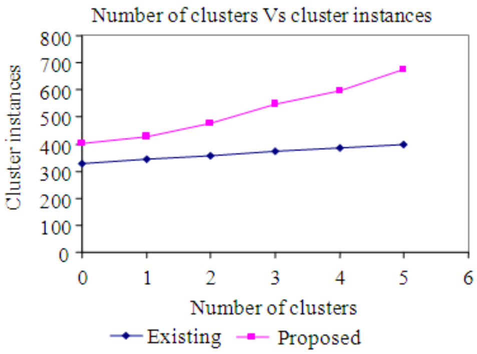

Fig. 3: Fuzzy K-Means cluster instances with number of clusters

The considered data sets for these experiments are synthetically generated according to the normal distribution. Figure 3 shows Cluster instances with the number of clusters. The execution time, as expected, is nearly quadratic with respect to the number of clusters but as c is usually a small integer, it creates no problem.

DISCUSSION

By using Weka the performance of educational institution quality factors has been evaluated. The system model has enabled instructors to efficiently create and distribute a wide variety of educational materials, assignments, assessments. These include numerous types of formative conceptual and algorithmic exercises for which prompt feedback and assistance can be provided to students as they work on assigned tasks. This process allows rapid interpretation of such data in identifying students' misconceptions and other areas of difficulty, so that concurrent or timely corrective action can be taken. This information also facilitates detailed studies of the educational resources used and can lead to redesign of both the materials and the course.

Since the output is the result of decision tree training modeling, therefore we require interpreting and generating explanation, which is understandable by humanity. Therefore the obtained decision tree is translated into rules. Explain one interesting rules among the various rules obtained. This rule has a purity of 55.6%. It represents the 22% (841students) of the total number of students. This rule is important because it provides new information. From the total number of students (841 students), 55.6% (468) are classified as “Successful”. The other students (44.4%, 373 students) are classified as “Unsuccessful”.

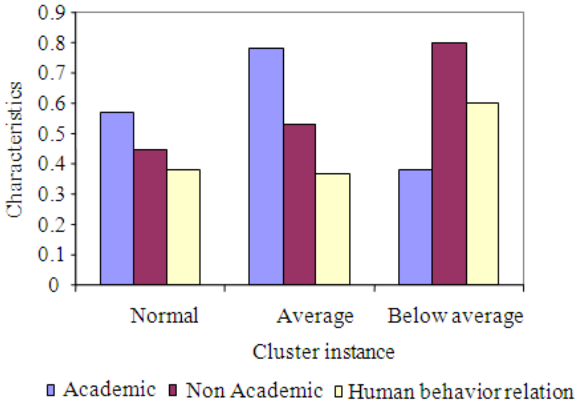

Fig. 4: Cluster instance Vs Institutional characteristics

Table 2: Decision support for the institutional performance metrics

Academic	Non-academic	Human behavior of Relation	Decision
0.57	0.45	0.38	Normal
0.78	0.53	0.37	Average
0.38	0.8	0.6	Below average

The Table 2 depicted below gives a overview of our proposed decision tree obtained in our experimentation on the institutional data set. The attributes selected using cluster analysis is feed into in the decision tree model and its resultant outcome is depicted in Fig. 4.

CONCLUSION

The proposed work of Fuzzy Modeled K-Cluster Quality Mining, evaluates the quality of clusters formed in the process of mining hidden data of education data sets. The validity is verified with cluster object cohesiveness and its precision value. The cluster validation is used to evaluate the quality index of the data set along with decision tree algorithm which presents the usability of Fuzzy k-Means efficiently. The decision tree obtained from clusters how close the results to the real partitions of the data set are. The proposal work defined the validity index, for assessing the results of clustering fuzzy k-means. The index is optimized for Institutional data sets that include compact and well-separated clusters. The compactness of the data set is measured by the intra-cluster density whereas the separation by the density between clusters.

REFERENCES

Al-Zoubi, M.B., A. Hudaib, A. Huneiti and B. Hammo, 2008. New efficient strategy to accelerate k-means clustering algorithm. Am. J. Applied Sci., 5: 1247-1250. DOI: 10.3844/ajassp.2008.1247.1250

Amirzadeh, V., M. Mashinchi and M.A. Yaghoobi, 2088. Construction of control charts using fuzzy multinomial quality. J. Math. Stat., 4: 26-31. DOI: 10.3844/jmssp.2008.26.31

Cheong, C.W., L.H. Jie, M.C. Meng and A.L.H. Lan, 2008. Design and development of decision making system using fuzzy analytic hierarchy process. Am. J. Applied Sci., 5: 783-787. DOI: 10.3844/ajassp.2008.783.787

Delavari, N., M.R. Beikzadeh and S. Phon-Ammuaisuk, 2005. Application of enhanced analysis model for data mining processes in higher educational system. Proceedings of the 6th International Conference on Information Technology Based Higher Education and Training, Jul. 7-9, IEEE Xplore Press, pp: 1-6. DOI: 10.1109/ITHET.2005.1560303

Feldman, R., Y. Regev, E. Hurvitz and M. Finkelstein-Landau, 2003. Mining the biomedical literature using semantic analysis and natural language processing techniques. BIOSILICO, 1: 69-80. DOI: 10.1016/S1478-5382(03)02330-8

Jaber, K., N.A. Rashid and R. Abdullah, 2009. The parallel maximal cliques algorithm for protein sequence clustering. Am. J. Applied Sci., 6: 1368-1372. DOI: 10.3844/ajassp.2009.1368.1372

Karakosta, C., H. Doukas and J. Psarras, 2008. A decision support approach for the sustainable transfer of energy technologies under the kyoto protocol. Am. J. Applied Sci., 5: 1720-1729. DOI: 10.3844/ajassp.2008.1720.1729

Thongwan, T., A. Kangrang and S. Homwuttiwong, 2011. An estimation of rainfall using fuzzy set-genetic algorithms model. Am. J. Eng. Applied Sci., 4: 77-81. DOI: 10.3844/ajeassp.2011.77.81

Xie, X.L. and G. Beni, 1991. A validity measure for fuzzy clustering. IEEE Trans. Patt. Anal. Mach. Intell., 13: 841-847. DOI: 10.1109/34.85677