Efficient Quality Assessment Technique with Integrated Cluster Validation and Decision Trees

S.Prakash Kumar
Senior Lecturer cum PhD Research Scholar
Department of Computer Applications
Erode Sengunthar Engineering College
Thudupathi, Erode-57, India

Dr.K.S.Ramaswami
Assistant Professor & Head (Regular)
Coimbatore Institute of Technology
Civil Aerodrome Post
Coimbatore-641 014,
Tamilnadu, India

ABSTRACT
Clustering becomes a key technique in analyzing quality assessment in most of the recent research works. The partitioned clustering techniques used in previous work utilize attributes of objects to form cluster. The cluster numbers were initialized, which reduces cluster quality in terms of cluster object aggregation and appropriation. The work presented an efficient quality assessment technique comprising of two parts i.e., fuzzy k-means cluster validation scheme and decision tree model. The Fuzzy k-means cluster validation scheme improves recall and precision measure of automatically labeling cluster objects. The decision tree model evaluates labeled cluster object and decides on the appropriation of attributes to its cluster validity index. The cluster quality index is measured in terms of number of clusters, number of objects in each cluster, cluster object cohesiveness, precision and recall values. Cluster validates focus on quality metrics of the institution data set features experimented with real and synthetic data sets. The results of quality indexed fuzzy k-means shows better cluster validation compared to that of traditional k-family algorithm. The experimental results of cluster validation scheme and decision tree confirm the reliability of quality validity index which performs better than other traditional k-family clusters.

Key words – Cluster Validation, Fuzzy K-Means, Quality Assessment

1. INTRODUCTION
The ability of data mining in improving the quality of educational processes assessment was recently done with clustering techniques. DM_EDU presented in [1, 2] used for the application of data mining in educational system. The methodology is based on CRISP-DM methodology CRoss Industry Standard Process for Data Mining [3].

1.1 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 [20]. 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 three approaches to investigate cluster validity [20]. The first is based on external criteria. This implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set. The second approach is based on internal criteria. We may evaluate the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves (e.g., proximity matrix). The third approach of clustering validity is based on relative criteria. Here the basic idea is the evaluation of a clustering structure by comparing it with other clustering schemes, resulting by the same algorithm but with different parameter values.

A number of validity indices have been defined and proposed in the literature for each of above approaches [20]. A cluster validity index for crisp clustering is proposed in [17], attempts to identify compact and well-separated clusters. Other validity indices for crisp clustering have been proposed in [16] and [19]. 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 [21]. In [18], an evaluation study of thirty validity indices proposed in the literature is presented.

1.2 Decision Tree in Institution Quality Assessment
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 [4, 5, 6]. Prior to the great usages that this technology brings into many application areas such as biomedical and DNA analysis [5, 7, 8], retail industry and marketing [5, 9], telecommunications [5, 10], web Mining [11], computer auditing [12], banking [5], fraud detection [10], financial industry [5] and medicine [13, 14], it recently has also been an interesting area of research in educational domain [15].

Nowadays the important challenge that education institutions face is reaching a stage to facilitate the universities in having more efficient, effective and accurate educational processes. Data mining is considered as the most suited technology appropriate in giving additional insight into the lecturer, student, alumni, manager, and other educational staff behavior and acting as an active automated assistant in helping them to make better decisions on their educational activities. Lack of deep and enough implicit knowledge in educational system may prevent system management to achieve their quality objectives. Data mining technology can help in bridging this knowledge gaps in educational system. Therefore the hidden patterns, association and anomalies, which are discovered by some data mining techniques, can be used to

improve the effectiveness, efficiency and the speed of the processes. As a result, this improvement may bring a lot of advantages to the educational system such as maximizing educational system efficiency, decreasing student's drop-out rate, increasing student's promotion rate, increasing student's retention rate, increasing student's transition rate, increasing educational improvement ratio, increasing student's success, increasing student's learning outcome, and reducing the cost of system processes. In order to achieve the above quality improvement, we need a data mining system that can provide the needed knowledge and insights for the decision makers in the educational system. The proposal in this paper presented an improved quality assessment model integrating fuzzy k-means and decision tree on the cluster formation. The enhanced version fuzzy k-means cluster along with decision tree improves performance of quality assessment in terms of precision and recall values. The quality improvement of decision-making is achieved through iterative fuzzy k-mean cluster. (i.e., student, faculty, and resource assessment under evaluation process)

2. INSTITUTIONAL QUALITY ASSESSMENT MODEL

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.

On the second process of predictive mining the new enhanced processes that data mining brings to educational system is by enhancing Student Examination sub process under Examination main process. Using fuzzy k-means applied on the set of student examination record grade and the lecturer performance and the academic activities, we can associate the exam level with lecturer class performance. The knowledge that can be extracted from this process would be the relation of exam level with lecturer performance and his/her academic activities in the class. The output presents how the performance of the lecturers in the class would vary with student exam level, or what activities the lecture should have done to increase the exam level grade with highest cluster object cohesiveness and cluster purity.

2.1 *Cluster Formation on Institutional parameters*

An 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. A small excerpt of the assignment performance in an introductory physics course, students in the rows, problems in the columns, each character representing one problem for one student are taken as sample piece. This view is particularly useful ahead of the problem deadline, where columns visualization with a large number of dots or blank spaces indicate problems that the students have difficulties with.

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 is created which includes all computed information from all students, sorted according to the problem order. In this step, the integrated fuzzy k-means and decision tree model has provided the following statistical information:

i) **Number of Students**: Total number of students who take a look at the problem. (Let number of students is equal to *n*)

ii). **Tries:** Total number of submissions to solve the problem

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

where *Xi* denote a student try.

iii) **Mod:** Mode, maximum number of submissions for solving the problem.

iv) **Mean:** Average number of the submissions.

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

v) **YES:** Number of students solved the problem correctly.

vi) **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.

vii) **%Wrng:** Percentage of students tried to solve the problem but still incorrect.

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

viii) **S.D.:** *Standard Deviation* of the students' submissions.

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

2.1.1 Fuzzy K-Means Cluster

Fuzzy K-Means (also called Fuzzy C-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.

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. For each iteration, the cluster output is stored in a directory cluster-N. The code has set number of reduce tasks equal to number of map tasks. Fuzzy KMeans iterates over input points and cluster points for specified number of iterations or until it is converged. During every iteration i, a new cluster-I directory is created which contains the modified cluster centers obtained during Fuzzy KMeans iteration. This will be feeded as input clusters in the next iteration. Once Fuzzy KMeans is run for specified number of iterations or until it is converged, a map task is run to output the point and the cluster membership to each cluster pair as final output to a directory named points.

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 propotional to the distance. Distance is computed using student assignment submission date, grade, etc as distance measure. Output key is encoded cluster. Output value is the probability value at input point. Fuzzy K-Means Combiner receives all key-value pairs from the mapper and produces partial sums of the cluster membership probability 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 summons all 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 the output. Output key is the encoded cluster identifiers.

2.1.2 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.

$$Inter\_dens(c) = \sum_{i=1}^{c} \sum_{\substack{j=1 \\ i \neq j}}^{c} \left(\frac{d\left(clos_{rep_i}, clos_{rep_j}\right)}{stdev_i + stdev_j}, density\left(u_{ij}\right) \right), c > 1, c \neq n$$

The criteria widely accepted for partitioning a data set into a number of clusters are separation of the clusters, and their compactness. Thus these criteria are obviously good candidates for checking the validity of clustering results. 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:

Density U_{ij} C>1,C≠n

$$Sep(c) = \frac{\sum_{i=1}^{c} \sum_{\substack{j=1 \\ i \neq j}}^{c} \min\{d(clos\_rep_i, clos\_rep_j)\}}{1 + Inter\_dens(c)}, c > 1$$

where clos_repi , clos_repj are the closest representative points between clusters i and j and n the number of points in a data set. Also, uij is the middle point of the line segment defined by the closest clusters' representatives clos_repi , clos_repj. The term density(uij) is defined as

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

where clos_repi, clos_repj are the closest representative points between cluster ci and cj 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 uij. The neighborhood of a data point, uij, is defined to be a hyper-sphere with center uij and radius the average standard deviation of the clusters between which we estimate the density. Also, the function
f(x, uij) is defined as:

$$f(x, u_{ij}) = \begin{cases} 0, & if\ d(x, u_{ij}) > (stdev_t + stdev_j)/2 \\ 1, & otherwise \end{cases}$$

It is obvious that a point belongs in the neighborhood of uij if its distance from uij 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 where clos_repi , clos_repj are the closest representative points between clusters ci and cj.

2.1.3 Institutional Parameters

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

a) Types of Informational Measures

i) Student profile
ii) Faculty profile
iii) Curriculum
iv) Outcome profile
v) Learning path ways
vi) Infrastructure facilities

b) Analytical Evaluation

i) Analysis of student involvement and engagement
ii) Staff Performance relating to student results and career
iii) Resource Facility relation to easy and effective knowledge acquisition
iv) Quality assessment in terms of faculty performance, student outcome, and resource availability
v) 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. The institution has adequate number of faculty with the appropriate expertise and competence to teach the courses offered.

The institution makes effective use of information and communications technology to support student learning and to manage its academic affairs and other programs and activities. The institution has a viable, sustainable and appropriate income generating strategy to support its development plans. 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.

2.2 Decision Tree for Predictive Table

Enhance the "Student assessment" sub-process under "Evaluation" main process. Using classification techniques like decision tree applied on the set of student and lecturer's academic and personal information, in an specific course, we are enable to classify students into various groups of successful and unsuccessful students. Therefore the knowledge that can be extracted from this process is the patterns of previously successful and unsuccessful students. By identifying these students known, we are able to decide which type of students are more successful than others and provide academic help for those who are less likely to be successful.

The reason of selecting student assessment sub-process is that each single student is considered as an asset in educational system, and educational domain has to put a great attempt in providing the highest investment on their talents, abilities, efforts, and interests while offering courses, therefore our concentration is on how to best assess student to improve their achievement in various courses. Based on the clustered data, the modeling functions classification is achieved through decision tree model. It is used to find the relationship between a specific variable (e.g., student race), the target variable (eg., student success status) and other variables (e.g., student grade, student number of failure in perquisite course) among the data. The outcome can be used to predict the correct class label to previously unseen and unlabeled student objects. (Fig 1) The algorithm applied on the set of data is decision tree therefore the result is presented in terms of a binary tree. There are many reasons of using decision tree in this application since it is relatively fast, it can be converted to simple and easy classification rule, it can be converted to SQL queries for accessing database, and it obtains similar or better accuracy in compassion with other classification techniques.

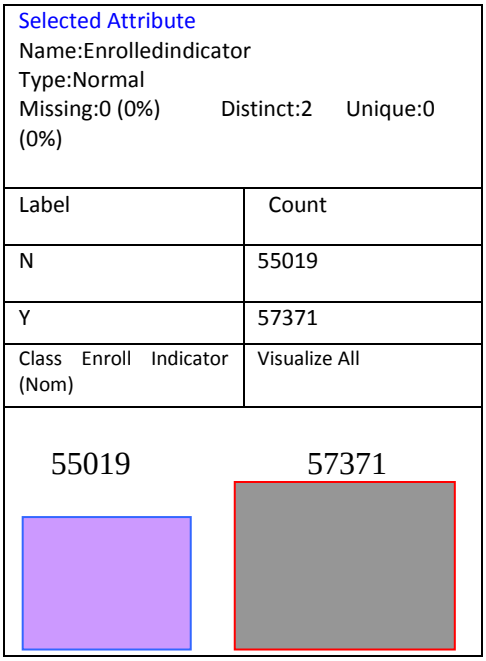

Fig 1: Student Enrollment decision indicator

3. EXPERIMENTAL RESULTS ON INTEGRATED FUZZY K-MEANS CLUSTER AND DECISION TREE

The experimentation conducted on the student assessment and the faculty

Table 1: Cluster formation of faculty associated student exam performance

Cluster	Students	Exams	Problems (each exam)	Pass	Attempts	Fail
Course	256	12	25	60	20	20
Faculty	240	10	20	65	25	10

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. A subjective number of four clusters were set for the model. Two distinctive clusters from the experimental resultant for display are

The characters between the course and faculty clusters are noticeable in student performance with respect to the faculty influence and course influence of student interest.(shown in Table 1) The percentage of pass and fail indicates that the overall cluster centroid of the faculty influence have produced better result for the students. An exercise like this is indeed valuable to assist researchers to evaluate the supervised modeling techniques and eventual decisions. The factor analysis reduces unwanted features on decision making process to identify the institutional quality.

3.1 Clustering Student Performance

The institution operates effective arrangements to direct scholarships and study grants on merit to support the most able students on programs that develop competencies needed to support the economy. The institution is effective in recruiting, admitting, supporting, and graduating students, including those from indigenous groups, the handicapped, low level income classes, foreign students, and other special groups. The institution has programs for student services, to support the non-academic needs of the students.

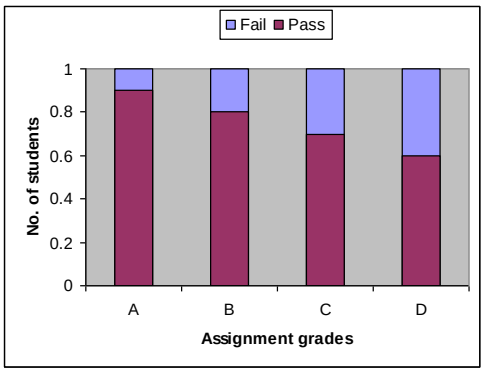

Figure 2: Student performance on assignment grades

The students are distributed based on the assignment grades obtained with the exam conducted. Graduation is becoming highly important in displaying the status as well as the quality of any educational institution. The resultant outcome shown in Figure 2 indicates the student cluster formation based on assignment grades depicts their quality percentage.

3.2 Cluster Quality on faculty performance

The institution sets the objectives and learning outcomes of its programs at appropriate levels by measuring faculty performance. The cluster quality analysis evaluates faculty performance in terms of student pass / fail percentage Institution needs effective arrangements for monitoring the effectiveness of its programs.

The institution takes effective action to address weaknesses, build on strengths, and to enhance performance by the dissemination of good practice. The quality of teaching evaluation made with respect to the student performance in terms of relevant faculty provides the efficiency level of the faculty. Figure 3 depicts the data samples of students undergoing exam associated to faculty

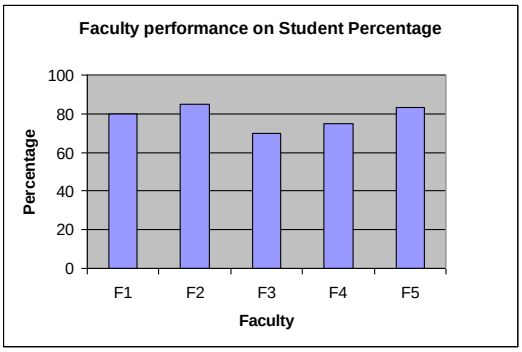

Figure 3: Faculty Student Exam Percentage

handling subjects. In its centers of excellence or of development the institution sustains consistently high levels of excellence in teaching, research and extension services.

3.3 Fuzzy K-means cluster Validation

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 setdimension); 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 $O(n)$.

Figure 4 and Figure 5 show graphical representation of experimental results conducted to identify execution time of fuzzy k-means cluster approach on faculty and student performance listing. The data sets for these experiments are synthetically generated from sample data obtained from local institutional body with normal distribution. Figure 4 shows the execution time as function of the number of clusters. The execution time is increasing as the

number of cluster keep on increasing with more number of data objects. As the increase in time is nominal it does not affect cluster validity.

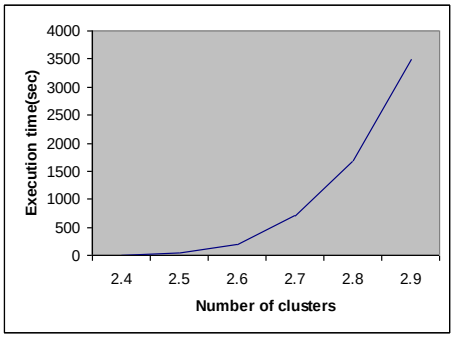

Figure 4: Fuzzy K-Means cluster execution time with number of cluster

The execution time calculated for cluster object tuples are almost linear as shown in Figure 5 which is presented graphically for fuzzy k-means on institutional quality metrics such as student and faculty performance. The proposal measured the execution time for data sets with higher dimensionality (two, four and six dimensions). Even then our proposed cluster scheme shows the linear outcome on the performance of execution time towards the cluster formation.

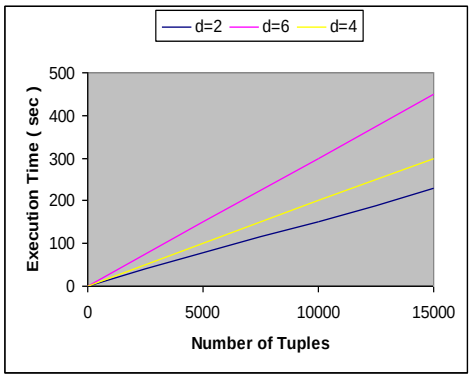

Figure 5: Fuzzy K-Means cluster execution time with number of points.

3.4 Decision Support Tree on the clustered data

The integrated fuzzy k-means and decision tree model has enabled to efficiently analyze the quality of institution based on educational materials, assignments, assessments, etc. 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. The decision on misconception is made decisively in concurrent or timely corrective action to be taken. This information also facilitates detailed studies of the educational resources used and lead to redesign of both the materials and the course.

The resultant output of decision tree training model interprets and generates explanation 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.

Table 2: Decision support for the Institutional performance metrics

S.No	Academic	Non-Academic	Human behaviour relation	Decision
1	0.8	0.6	0.8	Normal
2	0.7	0.4	0.9	Normal
3	0.5	0.4	0.4	Average
4	0.4	0.3	0.7	Below Average
5	0.3	0.4	0.9	Below Average
6	0.4	0.5	0.5	Below Average
7	0.8	0.6	0.8	Normal
8	0.8	0.6	0.8	Normal
9	0.8	0.3	0.8	Normal
10	0.3	0.2	0.4	Below Average
11	0.7	0.7	0.7	Normal

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".

Table 2 shown below gives the overview of our proposed decision tree obtained from 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 shown in Figure 6 graphically. The size of the decision tree with the number of leaves (i.e., faculty performance, student results and resource availability) on the performance of institutional quality indicates that the cluster object size have bigger impact on the quality assessment. In our work the quality assessment is effective to larger cluster objects with increased volume of data attributes.

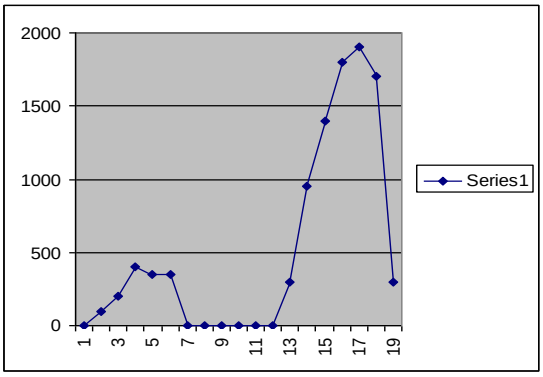

Figure 6: Size of Trees Vs Number of leaves

4. CONCLUSION

The quality assessment technique presented in this work integrated the fuzzy k-means algorithm with decision tree technique to provide an efficient quality analysis of Institutional data sets. The cluster validation scheme 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 evaluated with quality index of the institutional data set obtained from decision tree algorithm.

The decision tree is derived to assess institutional quality by utilizing intrinsic attributes of the institutional data. The validity index is used 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. The experimental result shows that integrated fuzzy kmeans with decision tree shows better quality assessment (nearly 22#) compared to traditional k-family clustering techniques.

5. REFERENCES

[1] Delavari N, Beikzadeh M. R, Shirazi M. R. A., "A New Model for Using Data Mining in Educational System", 5th International Conference on Information Technology based Education and Training: ITEHT '04, Istanbul, Turkey, 31st May-2nd Jun 2004.

[2] Delavari N, Beikzadeh M. R, "A New Analysis Model for Data Mining Processes in Educational Systems", MMU International Symposium on Information and Communications Technologies 2004 in conjuction with the 5th National Conference on Telecommunication Technology 2004, Putrajaya, Malaysia, 7th- 8th October 2004.

[3] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R, CRISP-DM 1.0: Step-by-step data mining guide, 2000

[4] Two Crows Corporation, "Introduction to Data Mining and Knowledge Discovery", TwoCrows Corporation, Third Edition, U.S.A, 1999.

[5] Han J, Kamber M, "Data Mining: Concepts and Techniques", Simon Fraser University, Morgan Kaufmann publishers, ISBN 1-55860-489-8. 2001.

[6] Chen M .S, Han J, Yu P. S, "Data Mining: An Overview from a Database Perspective". IEEE Transaction on Knowledge and Data Engineering, 1996.

[7] Han J, "How can Data Mining Help Bio-Data Analysis". BIOKDD02: Workshop on data mining in Bioinformatics, 2002.

[8] Feldman R, "Mining the Biomedical Literature using Semantic Analysis and Neural Language Processing Techniques, a link analysis approaches". ClearForest Corporation, New York, 2003.

[9] Edelstein H, "Building Profitable Customer Relationships with Data Mining", Two Crows Corporation, SPSS white paper-executive briefing, 2000.

[10] Chang W. H. T, Lee Y. H, " Telecommunications Data Mining for Target Marketing," Journal of Computers, Vol. 12, No. 4, December 2000, pp.60-74.

[11] Mobasher B, Jain N, Han E, Srivastava J, "Web Mining: Pattern Discovery from World Wide Web Transactions", Technical Report TR96-050, Department of Computer Science, University of Minnesota, 1996.

[12] The Y. W, Mustaffa K. M, Zaitun A. B, Lee, "Data Mining In Computer Auditing". Informing Science. Cork, Ireland June 19-21, 2002.

[13] Baylis P, "Better Health Care with Data Mining", SPSS White Paper, UK, 1999.

[14] Brossette S. E, Sprague A. P, Hardin J. M, Waites K. B, Jones W. T, Moser S.A, "Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance", Journal of the American Medical Informatics Association (JAMIA), vol. 5: 1998, pp.373-381.

[15] Luan J, "Data mining and Knowledge Management, A System Analysis for Establishing a Tiered Knowledge Management Model (TKMM)", Proceedings of Air Forum, Toronto, Canada. 2001.

[16] Rajesh N. Dave. "Validating fuzzy partitions obtained through c-shells clustering", Pattern Recognition Letters, Vol .17, pp613-623, 1996

[17] J. C. Dunn. "Well separated clusters and optimal fuzzy partitions", J. Cybern. Vol.4, pp. 95- 104, 1974

[18] Milligan, G.W. and Cooper, M.C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set", Psychometrika, 50, 159-179.

[19] Milligan G. W., Soon S.C., Sokol L. M. "The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 5, pp. 40-47, 1983

[20] S. Theodoridis, K. Koutroubas. Pattern recognition, Academic Press, 1999

[21] Xunali Lisa Xie, Genardo Beni. "A Validity measure for Fuzzy Clustering", IEEE Transactions on Pattern Analysis and machine Intelligence, Vol13, No4, August 1991.