ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st&22ndMarch, Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Intellectual Performance Analysis of Students by Using Data Mining Techniques

J.K. Jothi Kalpana, K. Venkatalakshmi

Dept of Computer Science and Engineering, V.R.S. College of Engineering and Technoloy, Arasur, Tamil Nadu, India.

University College of Engineering, Tindivanam, Tamil Nadu, India.

Abstract– Education Data Mining concerns the prediction of school failures in different levels such as primary, secondary and higher level. This paper intends to analysis the students' performance in different categories of measurements. In this analysis categorize the college student's academic performance for Villupuram district. Based on the clustering methods such as centroid based, distribution based and density based clustering. Cluster includes groups with small distance among the cluster members. The performance of student's multi-level of optimization formulated by using clustering. In centroid based clustering, clusters are represented by a central vector. The number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem. The clustering model most closely related to statistics is based on distribution model. Experiments attempts to improve the accuracy by using the method of Gaussian mixture model. The data set is modeled with a fixed number of Gaussian distribution that is initialized randomly and the parameters are iteratively optimized to fit better to the data set. The density based clustering method is a linkage based clustering. The range parameter ε produces a hierarchical result related to that of linkage clustering. Clustering can be represents in a large range of classifications and applications. K-means algorithm categorizes the large dataset. In this analysis use genetically improved particle swarm optimization algorithm to model the students level. The GAI-PSO algorithm searches the solution space to find the optimal result. The processing of refining use the k-means algorithm.

Keyword– Centroid based, Distribution based, Density based Cluster, K-means algorithm, Gaussian distribution.

I.INTRODUCTION

Recent years there are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining (EDM), concerned with developing methods that extract knowledge from data come from the educational context. The data can be collected from historical and operational data reside in the
databases of educational institutes. The student data can be an academic. Also it can be collected from e-learning systems which have a large amount of information used by most institutes.

In earlier research of Educational Data Mining predicting school failure in different educational level such as primary, secondary and higher level [2]. This way of grasping knowledge in databases, called Educational Data Mining (EDM). The analysis of this educational
mining uses many approaches and techniques such as decision tree, Rule induction, Neural network, K-nearest neighbor and Naïve Bayesian.

The main objective of higher education institutes is to provide quality education to its students and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main attributes that may affect the students' performance [5]. The discovered knowledge can be used to offer a helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve students' academic performance and trim down failure rate, to better understand students' behavior, to assist instructors, to improve teaching and many other benefits [2].

Improved educational data mining uses many techniques such as centroid based, distribution based and density based clustering. Cluster includes groups

with small distance among the cluster members. By using these techniques, many kinds of knowledge can be discovered such as K-means and Gaussian.

This paper investigates the improved educational domain of data mining analysis the graduate students data collected from the college of Engineering and Technology Villupuram. The data include five years period [2008-2013]. It showed what kind of data could be collected, how could we preprocess the data, how to apply data mining methods on the data, and finally how can we benefited from the discovered knowledge. There are many kinds of knowledge can be discovered from the data. In this work we investigated the most common ones which are centroid based, distribution based and density based clustering. Cluster includes groups with small distance among the cluster members. The Matlab software is used for applying the methods on the Engineering student's data set.

Through this discovered knowledge, we need to provide a college management with a helpful and constructive recommendation to overcome the problem of low grades of graduate students, and to improve students' academic performance.

This paper takes into consideration of the official approval from the college of Engineering and Technology - Villupuram was obtained to have an access to the related databases for the sole use of analysis and knowledge discovery purposes. To achieve result, all and individual person data are extracted from the database before applying the data mining methods.

The rest of this paper is organized as follows: Section 2 presents related works in educational data mining. Section 3 describes the data set and the preparation and processing methods performed. Section 4 reports our experiments about applying data mining techniques on the educational data. Finally we conclude this paper with a summary and an outlook for future work in the below Section.

II. ASSOCIATED WORKS

Although, using data mining in higher education is a recent research field, there are many works in this area. That is because of its potentials to educational institutes. The Educational Data Mining (EDM) is a promising area of research and it has a specific requirements not presented in other domains. Thus, work should be oriented towards improved educational domain of data mining. In the existing system, predicting the academic failure of students in different levels such as primary, secondary and higher. The methods to analysis these levels by Gathering students' data, Pre-processing, Data mining and Interpretation [2]. For example the **existing system presents**, the numerical values of the scores obtained by students in each subject were changed to categorical values in the following way:

TABLE I
STUDENTS' DATA

S.No.	Source Variable	Variable
1.	Specific survey	Classroom/group, number of students in group, attendance during morning/evening sessions, number of friends, number of hours spent studying daily, methods of study used, place normally used for studying, having one's own space for studying, resources for study, study habits, studying in group, parental encouragement for study, marital status, having any children, religion, having administrative sanctions, the type of degree selected, the influence on the degree selected, the type of personality, having a physical disability, suffering a critical illness, regular consumption of alcohol, smoking habits, family income level, having a scholarship, having a job, living with one' s parents, mother' s level of education, father' s level of education, number of brothers/sisters, position as the oldest/middle/youngest child, living in a large city, number of years living in the city, transport method used to go school, distance to the school, level of attendance during classes, level or boredom during classes, interest in the subjects, level of difficulty of the subjects, level of motivation, taking notes in class, methods of teaching, too heavy a demand of homework, quality of school infrastructure, having a personal tutor, level of teacher' s concern for the welfare of each student.
2.	Coeval	Age, sex, previous school, type of school, type of secondary school, Grade Point Average (GPA) in secondary school, mother' s occupation, father' s occupation, number of family members, limitations for doing exercises, frequency of

		exercises, time spent doing exercises, score obtained indifferent subjects, and average score in the EXAMINATION I.
3.	Department of school services	Score in Math 1, score in Physics 1, score in Social Science 1, score in Humanities 1, score in Writing and Reading 1, score in English 1, and score in Computer 1.

The educational data mining is used to analyze students' learning behavior. The goal of the study is to show how useful data mining can be used in higher education to improve students' performance. He used students' data from database course and collected all available data including academic records of students, course records. Then applied the improved data mining (IDM) techniques to discover many kinds of knowledge such as centroid based, distribution based and density based clustering. Cluster includes groups with small distance among the cluster members. Also this can clustered the student into groups using Centroid, and detected all similarities in the data mining analysis. Finally, this can prove how we can benefit from the discovered knowledge to improve the performance of student.

The data mining techniques, particularly classified to help in improving the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses. The extracted classification rules are based on the different data the extracted classification rules are studied and evaluated. It allows students to predict the final grade in a course under study.

The classification of data mining technique to evaluate students' performance, they used K-Means method for classification. The goal of their study is to extract knowledge that describes students' performance in end semester examination. They used students' data from the students' previous database including Attendance, Class test (Prior Class Assessment Test and Class Assessment Test), Seminar, and Assignment marks [2]. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising.

The mining analysis based on students' failed courses to identifies students' failure patterns. The goal of their study is to identify hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity students' performances. The extracted density rules reveal some hidden patterns of students' failed courses which could serve as a foundation stone for academic planners in making academic decisions and an aid in the curriculum re-structuring and modification with a view to improving students' performance and reducing failure rate [5].

The K-Means clustering algorithm as a data mining technique to predict students' learning activities in a students' database including class quizzes, mid and final exam and assignments. These correlated information will be conveyed to the class teacher before the conduction of final exam. This study helps the teachers to indicate the details of the targeted students' performance and reduce the failing ratio by taking appropriate steps at right time and improve the performance of students.

III. THE ENGINEERING STUDENTS PREPROCESSING

The CGPA (Cumulative Grade Point Average) attribute in the data set contains a large number of continuous values. So for efficient later processing, simplified data description and understanding for data mining results, we credited this attribute to categorical one. For example, we grouped all GPAs into five categorical segments; Excellent, Very good, Good, Average and Poor.

After applying the preprocessing and preparation methods, we try to analyze the data visually and figure out the distribution of values, specifically the grade of students. Figure 1 depicts the distribution of graduate students in period from 2008 to 2013 according to their grades, it is apparent from the figure that the average students present about 54% of the data set.

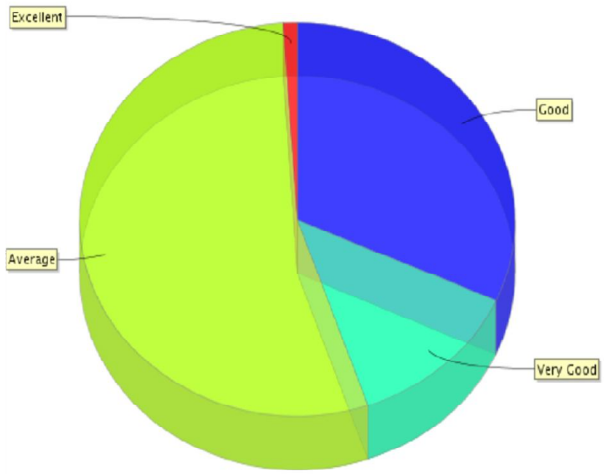


Fig.1 The distribution of engineering students according to their grades

IV.APPLICATIONOFDATAMINING TECHNIQUES TO ENGINEERING STUDENTS DATASET: RESULTS AND DISCUSSION

Before applying the data mining techniques on the data set, there should be a methodology that governs our work. Figure 2 depicts the work methodology used in this paper, which is based on the framework proposed. The methodology starts from the problem definition, then preprocessing which are discussed in the introduction and the data set and preprocessing sections, then we come to the data mining methods which are centroid based, distribution based and density based clustering. Cluster includes groups with small distance among the cluster members. Finally, the

knowledge representation processes the steaming of data and handle large datasets.

In this section, we describe the results of applying the data mining techniques to the data of our case study, for each of the four data mining tasks; Centroid, Distributed, Density based and Knowledge representation, and how we can benefit from the discovered knowledge.

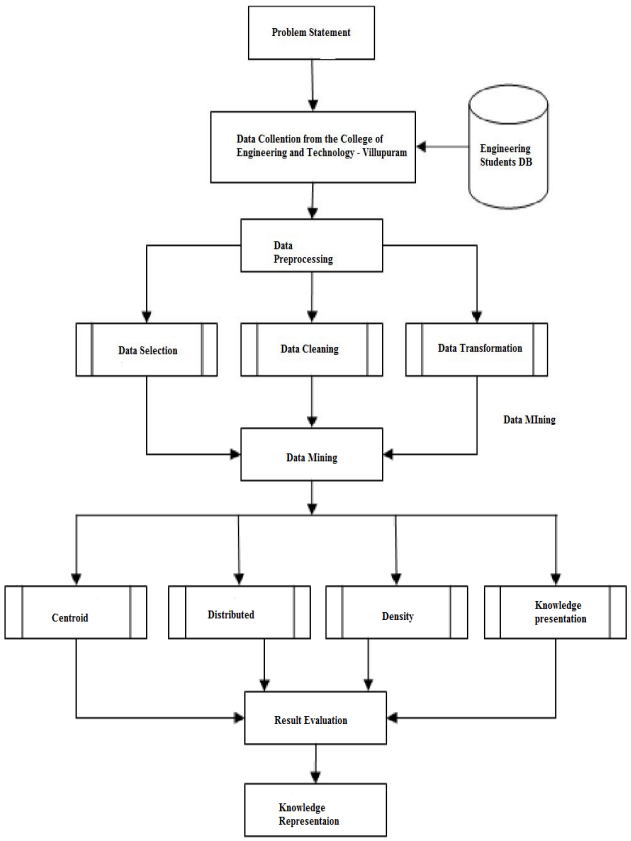


Fig. 2 Data Mining Work Methodology.

A. CENTROID BASED ALGORITHM

Let R^{dim} be the Euclidean space of dimension *dim;* $S \subseteq R^{dim}$ be a finite subset of data of size $N = |S|$; and $M = \{m_k | k=1,\ldots,K\}$, the set of parameters to be optimized. (The parameter set M consists of K centroids for K-Means, K centers for K-Harmonic Means, and K centers with co-variance matrices and mixing probabilities for EM.) We write the performance function and the parameter optimization step for this class of algorithms in terms of SS. The performance function is decomposed as follows:

1) Performance Function

$$F(S, M) = f0\ (\sum f1(x,M), \sum f2(x,M), \ldots\ldots, \sum fR(x,M)). \quad (1)$$

What is essential here is that f_0 depends only on the SS, represented by the sums, whereas the remaining f_i functions can be computed independently for each data point. The detailed form of f_i, $i=1,\ldots,R$, depend on the particular performance function considered. It will become clear when examples of K-Means, K-Harmonic Means and EM are given in later sections.

We write the center-based algorithm, which minimizes the value of the performance function over M, as an iterative algorithm in the form of Q SS ($I()$ stands for the iterative algorithm, and $\sum g_j$, $j=1,\ldots,Q$, stands for SS.):

$$M^{(u+1)} = I(\sum_{x \in S} g_1(x,M^{(u)}), \sum_{x \in S} g_2(x,M^{(u)}), \ldots\ldots, \sum_{x \in S} g_Q(x,M^{(u)})). \quad (2)$$

$M^{(u)}$ is the parameter vector after the u^{th} iteration. We are only interested in algorithms that converge: $M^{(u)} \to M$. The values of the parameters for the 0^{th} iteration, $M^{(0)}$, are by initialization. One method often used is to randomly initialize the parameters (centers, covariance matrices and/or mixing probabilities). There are many different ways of initializing the parameters for particular types of center-based algorithms in the literature [2]. The computation carried out here will be identical to the traditional, sequential equivalent. The set of quantities

$$Suff = \{\sum_{x \in S} f_r(x, M) \mid r = 1,\ldots\ldots,R\} \bigcup \{\sum_{x \in S} g_q(x,M) \mid q = 1,\ldots\ldots,Q\}. \quad (3)$$

$$Suff_l = \{\sum_{x \in D_l} f_r(x, M) \mid r = 1,\ldots\ldots,R\} \bigcup \{\sum_{x \in D_l} g_q(x, M \mid q = 1,\ldots\ldots,Q\}. \quad (4)$$

is called the global SS of the problem (1)+(2). As long as these quantities are available, the performance function and the new parameter values can be calculated and the algorithm can be carried out to the next iteration. We will show in Section 4 that K-Means, K-Harmonic Means and Expectation-Maximization clustering algorithms all belong to this class defined in (1)+(2).

2) Clustering Algorithm

Step 1: Initialization: Partition the data set and load the l^{th} partition to the memory of the l^{th} computing unit. Use any preferred algorithm to initialize the parameters, $\{m_k\}$, on the Integrator.
Step 2: Broadcast the integrated parameter values to all computing units.
Step 3: Compute at each unit independently the SS of the local data based on (4).
Step 4: Send SS from all units to the Integrator
Step 5: Sum up the SS from each unit to get the global SS, calculate the new parameter values based on the global SS, and evaluate the performance function. If the condition is not met goto step1 for the next iteration, else inform to stop all computing unit to stop.

B. Distributed Data Clustering

Data clustering is the task of partitioning a multivariate data set into groups maximizing intra-group similarity and inter-group dissimilarity. In a distributed environment, it is usually required that data objects are not transmitted between sites for efficiency and security reasons. An approach to clustering exploits the local maxima of a density estimate (Density Estimate) to search for connected regions which are populated by similar data objects [8]. In a scheme for distributed clustering based on DE has been proposed. Every

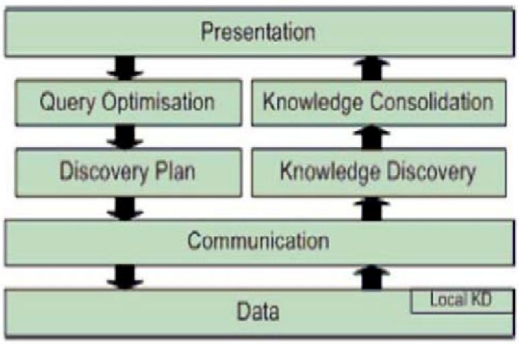


Fig. 3 Overview of DDC system

participating site computes a DE based on its local data only. Then, every site applies information theoretic regular multidimensional sampling to generate a finite, discrete, and approximate representation of the DE, consisting of its values at a finite number of equidistantly spaced locations. The samples computed by all sites are transmitted and summed (by location) outside the originating site, e.g., at a distinguished helper site. The resulting list of samples, which is an approximate representation of the true global DE, is transmitted to each participating site.

Every site executes a density-based clustering algorithm to cluster its local data with respect to the global DE, the values of which can be computed from the samples by means of a sampling series. Notice that a DE is not a band-limited function, therefore sampling produces aliasing errors, which increase as the number of samples decreases.

We propose to implement the approach by a society of agents. For example, in a real scenario all participating agents belong to different competing organizations, which agree to cooperate in order to achieve some common goal, without disclosing the contents of their data banks to each other. Each agent will negotiate with other agents to evaluate the advantages and risks which derive from participating to the distributed mining task. In particular, considerable security risks arise from the potential ability of the other agents to carry out inference attacks on density estimates. The resulting disclosure of sensitive information could be exploited as a competitive advantage by the organizations which own the malicious agents. Other aspects an agent has to evaluate in order to autonomously decide whether it should participate or not, include, but are not limited to, investigating a probabilistic model of trustworthiness of participating agents, the relation between trustworthiness and the topology of participating agents, and the probability of incurring coalition attacks.

C. Density Based Cluster

Distributed Clustering assumes that the objects to be clustered reside on different sites. Instead of transmitting all objects to a central site (also denoted as server) where we can apply standard clustering algorithms to analyze the data, the data are clustered independently on the different local sites (also denoted as clients). In a subsequent step, the central site tries to establish a global clustering based on the local models, i.e. the representatives [4]. This is a very difficult step as there might exist dependencies between objects located on different sites which are not taken into consideration by the creation of the local models. In contrast to a central clustering of the complete dataset, the central clustering of the local models can be carried out much faster.

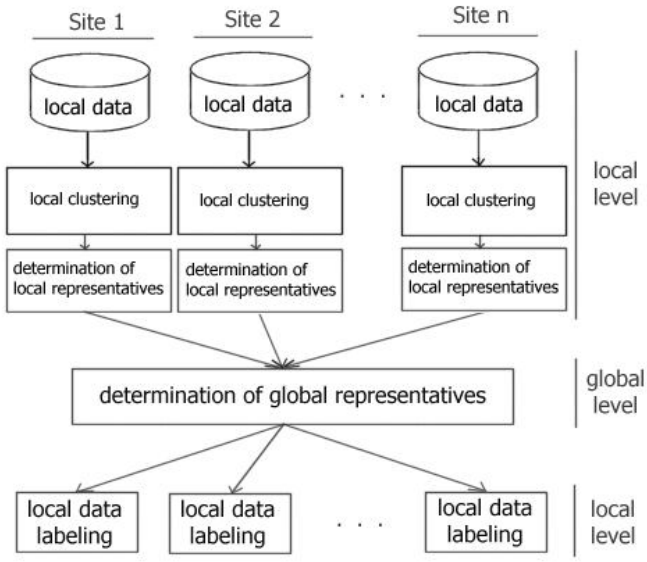


Fig. 4 Distributed Clustering

Distributed Clustering is carried out on two different levels, i.e. the local level and the global level. On the local level, all sites carry out a clustering independently from each other. After having completed the clustering, a local model is determined which should reflect an optimum trade-off between complexity and accuracy. Our proposed local models consist of a set of representatives for each locally found cluster. Each representative is a concrete object from the objects stored on the local site. Furthermore, we augment each representative with a suitable ε−range value. Thus, a representative is a good approximation for all objects residing on the corresponding local site which are contained in the ε−range around this representative. Next the local model is transferred to a central site, where the local models are merged in order to form a global model.

The global model is created by analyzing the local representatives. This analysis is similar to a new clustering of the representatives with suitable global clustering parameters. To each local representative a global cluster-identifier is assigned. This resulting global clustering is sent to all local sites. If a local object belongs to the ε-neighborhood of a global representative, the cluster-identifier from this representative is assigned to the local object. Thus, we can achieve that each site has the same information as if their data were clustered on a global site, together with the data of all the other sites.

To sum up, distributed clustering consists of four different steps:

• Local clustering
• Determination of a local model
• Determination of a global model, which is based on all local models
• Updating of all local models

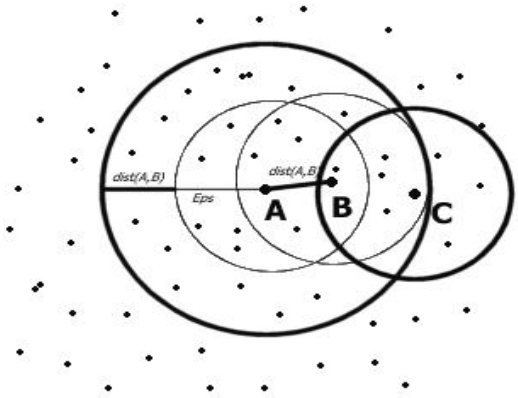


Fig. 5 REPScor: specific core points and specific ε−range

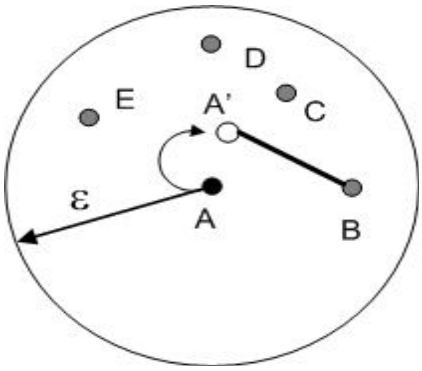


Fig. 6 REPk-*Means*: representatives by using k-means

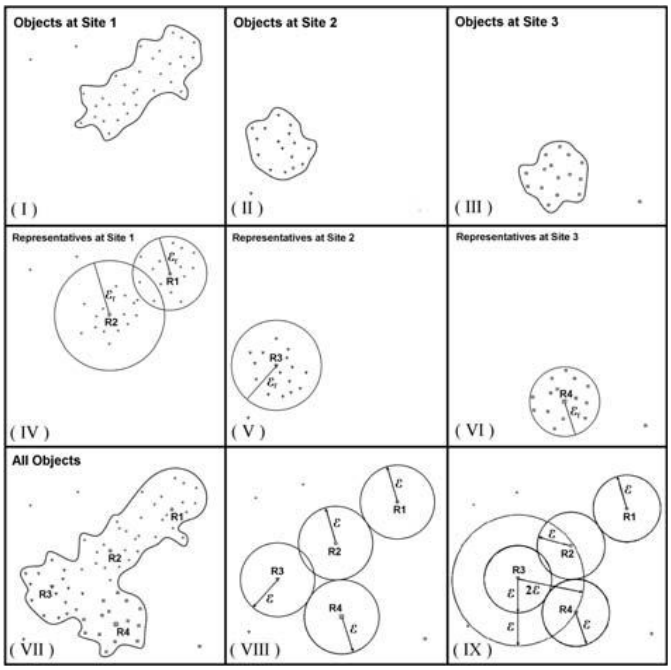


Fig. 7 Determination of a global model

TABLE II
COMPARISON REPORT

S.No.	Central Clustering	Distributed Clustering	Density based Clustering
1.	The mathematic of a class of clustering algorithms to reveal a straightfor ward parallel implement ation based on communica ting only a small amount of data.	The application of clustering algorithms to large spatial databases raises the minimal number of input parameters, discovery of clusters with arbitrary shape and efficiency on large databases.	A partitioning clustering algorithm which is bases on the density-based clustering algorithm.
2.	Highly efficient speed-up and scale-up for very large data sets.	Large spatial databases is very attractive when considering its non-parametric nature and	An enormous efficiency advantages compared to a central clustering carried out on all the

		its good quality for clusters of arbitrary shape.	quoted data.
3.	The ideas presented in this mining apply to iterative parameter estimation algorithms where the size of the SS (small set of data) is small relative to the data size.	The analysis in this mining is based on the assumption that the point inside of a cluster are uniformly distributed.	Based on the quality, new distributed clustering approach yields almost the same clustering quality as a central clustering on all data.

D. Knowledge Representation

1) K-means Algorithm

A partition clustering algorithm splits the data points into k partitions, where each partition represents a cluster. The partitioning is done based on certain objective function. One of the criterion functions is minimizing square error criterion which is computed as shown by formula:

$$E = \Sigma\Sigma \|p - \mu\|^2$$

Where p is the point in a cluster and μi is the centroid of the cluster. Each cluster must have at least one point and each point must be in one and only one cluster.

K-means is one of the most widely used partition-based clustering algorithms in practice [3]. It is simple, easy, understandable, scalable, and can be adapted to deal with streaming data and very large datasets. K-means algorithm divides a dataset X into k disjoint clusters based on the dissimilarities between data objects and cluster centroids. Let be the centroid of cluster Ci and the distances between Xj that belong to Ci and ¯is equal to d(Xj, ¯).

2) Pseudo code for K-means algorithm

Require: k ≥ 2 and t ≥ 1 {
1: Select initial cluster centroids
2: Repeat
3: For each point xj in a dataset do
4: For all do
5: Compute the dissimilarity
6: End for.
7: assign point xj to closest cluster Ci;
8: End for.
9: For all do
10: Update as the centroid of cluster Ci;
11: End For
12: Until convergence criterion is satisfied or the number of iterations exceeds a given limit t.

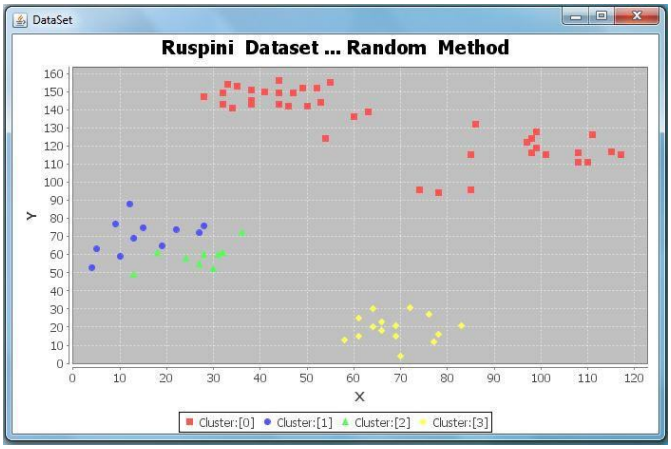


Fig. 8 Results of running the K-means with k=4and using 4 different starting points, each randomly chosen from the dataset

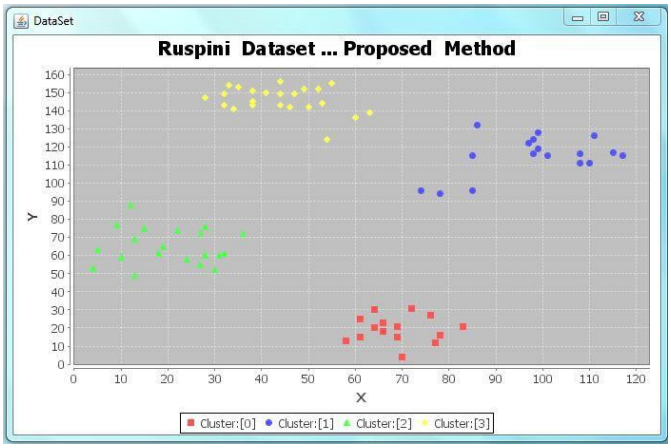


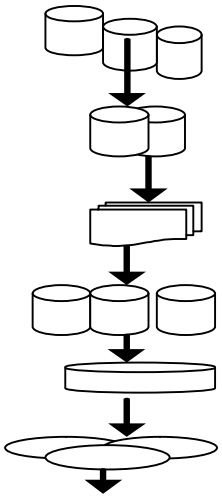
Fig. 9 Results of running the K-means with k=4 and using 4 different starting points, each chosen with the proposed method

3) Gaussian Processes

Gaussian processes (GPs) are a modeling mechanism with origins in spatial statistics, particularly rigging [Journal and Huijbregts, 1992]. In contrast to global approximation techniques such as least-squares fitting, GPs are local approximation techniques, akin to nearest neighbor procedures [7]. In contrast to function approximation techniques that place a prior on the form of the function, GP modeling techniques place a prior on the covariance structures underlying the data.

The basic idea in GPs is to model a given dataset as a realization of a stochastic process. Formally, a GP is a set of random variables any finite subset of which have a (multivariate) normal distribution [6]. For our purposes, we can think of these

variables as spatially distributed (scalar) response variables ti, one for each 2D location xi = [xi1, xi2] where we have collected a data sample. In our vector field analysis application, ti denotes the modeled response, i.e., the value of de Boor's function at xi. Given a dataset D = {xi, ti}, i = 1 . . . n, and a new data point xn+1, a GP can be used to model the posterior P(tn+1|D, xn+1) (which would also be a Gaussian). This is essentially what many Bayesian modeling techniques do (e.g., least squares approximation with normally distributed noise) but it is the specifics of how the posterior is modeled that make GPs distinct as a class of modeling techniques.



Extracted Knowledge

Fig. 9 Steps of Knowledge Extraction

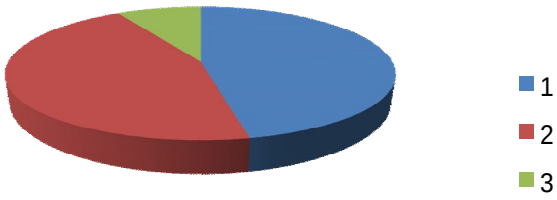


Fig. 10 Sample Percentages' of student CGPA

The benefit of these two methods is that it can predict low grades on time. For example the college management can predict Average students from the beginning and they may work on them to improve their performance before the graduation.

SimpleDistribution

```
Distribution model for label attribute Grade

Class Good (0.318)
5 distributions

Class Very good (0.130)
5 distributions

Class Excellent (0.010)
5 distributions

Class Average (0.542)
5 distributions
```

Fig. 11 The Distribution model for label attribute Grade

V. CONCLUSION AND FUTURE WORK

In this paper, we gave a case study in the educational data mining. It showed how useful data mining can be used in higher education particularly to improve graduate students' performance. We used students data collected from the college of Engineering and Technology in Villupuram. The data include five years period [2008-2013]. We applied data mining techniques to discover knowledge. Based on the clustering methods such as centroid based, distribution based and density based clustering. Cluster includes groups with small distance among the cluster members. Also we clustered the students into groups using K-Means clustering algorithm. Finally the Distance-based Approach and Density-Based Approach are used. Each one of these tasks can be used to improve the performance of graduate student.

Our future work include applying data mining techniques on an expanded data set with more distinctive attributes to get more accurate results.

REERENCES

[1] Bin Zhang, Meichu Hsu, George Forman, (2000) 'Accurate Recasting of Parameter Estimation Algorithms Using Sufficient Statistics for Efficient Parallel Speed-up – Demonstrated for Center-Based Data Clustering Algorithms.

[2] Carlos Marquez-Vera, Cristobal Romero Morales and Sebastian Ventura Soto, (2013) 'Predicting School Failure and Dropout by Using Mining Techniques', Vol. 8, No.1.

[3]Chunfei Zhang, Zhiyi Fang, (2013) 'An Improved K-means Clustering Algorithm', Journal of Information and Computational Science 10: pp. 193-199.

[4] Huan Wang, Yanwei Yu, Qin Wang and Yadong Wan, (2012) 'A Density –Based Clustering Structure Mining Algorithm for Data Streams'.

[5] Mohammed M. Abu Tair, Alaa M. EI-Halees, (2012) 'Mining Educational Data to Improve Students' Performance: A Case Study', International Journal of Information and Communication Technology Research, Vol. 2 No.2, pp. 140-146.

[6]NarenRamakrishnan, Chris Bailey-Kellogg, SathishTadepalli, and Varun Pandey 'Gaussian Processes for Active Data Mining of Spatial Aggregates'.

[7] Neha Aggarwal, Faridabadl, and Kiriti Aggarwal . (2012) 'A Mid- Point based K-mean Clustering Algorithm for Data Mining', International Journal on Computer Science and Engineering', Vol.4 No. Global Journal of Computer Science and Technology, vol. 10, no. 06, pp. 1174-1180.

[8] XiaoweiXu, Martin Ester, Han-Peter Kriegel, and Jorg Sander, (1998) 'A Distributed-Based Clustering Algorithm for Mining in Large Spatial Databases', International Conference on Data Engineering.