Abstract — Educational Data mining is a recent trend where data mining methods are experimented for the improvement of student performance in academics. The work describes the mining of higher education urban and rural students’ related attributes such as improvement, behavior, attitude and relationship. The data were collected from Dr.Ambedkar Government Arts College in terms of the mentioned attributes. The proposed work explored improvement Behavior Attitude Relationship Clustering (IBARC) Algorithm, which showed the improvement in students’ performance in terms of Improvement predicting good behavior, average attitude and good relationship with faculty members and Tutors. The Hierarchical clusters were grouped with related similarities and analysis was experimented using WEKA tool. The resulted analysis describes the input parameters were found optimal. These systems do not permit the closely monitoring of students’ academic activities, the evaluation of their academic success and the support of academic activities concerning with teaching and tutoring. This dissertation makes use of cluster analysis to segment students into groups according to their characteristics and also will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students. It also focuses on feedback analysis of teachers by students. Clustering is one of the basic techniques often used in analyzing data sets. The main goal of clustering is to partition students into homogeneous groups according to their characteristics and abilities. These applications can help both instructor and student to enhance the quality education. Improving students’ academic performance is not an easy task for the academic community of higher learning.

Keywords— Improvement, Behavior, Attitude, Relationship, Methods of Clustering, Urban and Rural Students’ Performance.

I. INTRODUCTION

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in [1]. There are increasing research interests in using data mining techniques in educational filed. This new emerging field, EDM, concerns with developing methods that discover knowledge from data originating from educational environments. Educational data mining techniques often differ from traditional data mining techniques, in explicitly exploiting the multiple levels of meaningful hierarchy in educational data.

1.1 EDM METHODS

Romero and Ventura [2007] categorize work in educational data mining into the following categories:

a) Statistics and visualization
b) Web mining
(i) Clustering, classification, and outlier detection
(ii) Association rule mining and sequential pattern mining
(iii) Text mining
Baker, which classifies work in educational data mining as follows:

a) Prediction
(i) Classification
(ii) Regression
(iii) Density estimation
b) Clustering
c) Relationship mining
(i) Association rule mining
(ii) Correlation mining
(iii) Sequential pattern mining
(iv) Causal data mining
d) Distillation of data for human judgment
e) Discovery with models
DATA MINING AND TECHNIQUES

Data Mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing knowledge sharing to make proactive, knowledge-driven decisions.

The most commonly used techniques in data mining are:

1. Artificial Neural Networks: this is a nonlinear predictive model that learns through training and resembles biological neural networks in structure.
2. Decision trees: tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.
3. Genetic Algorithms: They are optimization techniques that use process such as genetics combination, mutation, and natural selection in a design based on concepts of evolution.
4. Rule Induction: the extraction of useful if-then rules from data based on statistical significance.
5. Regression Methods: this tries to identify the best linear pattern in order to predict the value of one characteristic we are studying in relation to another.

Data mining does various types of tasks like,

Classification is a task of finding a function that maps records into one of several discrete classes. Clustering is a task of identifying groups of records that are similar between themselves but different from the rest of the data.

II. LITERATURE SURVEY

Zaïane (2001) suggests an application for data mining, using it to study on-line courses. This article proposes and evangelizes EDM’s usefulness, and in this fashion was highly influential to the formation of our community.

Zaïane (2002) and Tang and McCalla (2005) center around how educational data mining methods (specifically association rules, and clustering to support collaborative filtering) can support the development of more sensitive and effective e-learning systems.

Tang and McCalla report an instantiation of such a system, which integrates clustering and collaborative filtering to recommend content to students.

Baker, Corbett and Koedinger (2004) gives a case study on how educational data mining methods can be used to open new research areas, in this case the scientific study of gaming the system.

Merceron and Yacef (2003) and Romero et al. (2003), present tools that can be used to support educational data mining.

Beck and Woolf (2000) shows how educational data mining prediction methods can be used to develop student models. They use a variety of variables to predict whether a student will make a correct answer. This work has inspired a great deal of later educational data mining work – student modeling is a key theme in modern educational data mining, and the paradigm of testing EDM models’ ability to predict future correctness.


Rajendran et al. (2007) studied that are rural students inferior to urban students in their achievement scores in chemistry at college level and found that there was no significant difference among the achievement of boys as well as girls in the post test of chemistry; locality of students (urban/rural) had no influence on the achievement scores of students at college level.

Ojoawo (1989) studied the effects of differential distribution of resources on school performances in an examination and found that location of schools in Oyo state had significant effect on school academic performance and there was significant difference in the performance between the students of rural and urban schools.

Bookman (1996) studied academic adjustment in relation to scholastic achievement of secondary school pupils by taking a sample of 545 senior secondary school students and found that academic adjustment was significantly related to the scholastic performance; the scholastic performance and locality were unrelated; there was no difference among the subjects from urban, semi-urban and rural localities with regard to scholastic performance.

III DATA COLLECTION

The students’ data were collected as a dataset from a Department of computer science Dr Ambedkar Govt. Arts college. To predict the attributes such as improvement, behavior, attitude and student staff relationship interaction, the similar set of data were predicted and clustered under a tutor. Where, tutor is a teaching staff and they are responsible for the set of student under them. As it is related with the students’ cognitive intelligence, using tutoring system the data were collected for experiment.
The basic idea of clustering is to define the similarity between the distance, the distance that represents the data between the data to measure the similarity of the size of the data are classified, until all the data gathering is completed. The proposed work uses hierarchical clustering analysis using the parameters as improvement, behavior, attitude and relationship. Hierarchical clustering creates a hierarchy of clusters which may be represented in a tree structure called a dendrogram. Clustering process to construct a spanning tree, which contains the class hierarchy information, and the similarity between all classes and class, it generates a hierarchical nested class, and can be used for any feature type, high accuracy. But because each merger, the global compare all the similarity between the class and choose from the smallest two classes, so run slower, not suitable for large-scale data sets.

IBARC algorithm describes students’ improvement, Behavior, Attitude and Relationship with tutor and Staff. The parameters correlation can be experimented using matrix structure. The attribute behavior is termed as ($\alpha$), Attitude as ($\beta$) relationship ($\gamma$), and improvement as($y$).

The relation between the attributes is as shown below.

$$D = C_1 (\alpha + \beta + \gamma + y) \ldots \ldots C_i \ldots C_n (\alpha + \beta + \gamma + y), \text{ Where } \alpha, \beta, \gamma, y \text{ be } 0, 1, -1, 2 \ldots \ldots \ldots \ldots (1)$$

From the equation 1, the matrix of each clusters are derived.

IBARC ALGORITHM

1. Choose D as a dataset;
2. Data Collection, C = ($\alpha, \beta, \gamma, y$ ), of C1…n ($\alpha, \beta, \gamma, y$)1….n, where
D is to identify the clusters of optimal similarities relates to the attributes behavior, attitude and relationship.
3. IF C=1 AND ||D||=1 THEN
4. Assign that student in that clusters as a optimal improvement, behavior, attitude and relationship;
5. Randomly select a center point data, that exhibits the student performance when $\alpha, \beta, \gamma, y$ = 1 or -1.
6. Repeat to group a cluster until $\alpha, \beta, \gamma, y$ applies to 1 or -1.
7. If student $\alpha, \beta, \gamma, y$ = 0, then avoid the student from the cluster.

V. EXPERIMENT USING WEKA

The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. It is freely available software. It is portable & platform independent because it is fully implemented in the Java programming language and thus runs on almost any modern computing platform. Weka has several standard data mining tasks, data preprocessing, clustering, classification, association, visualization, and feature selection. The WEKA GUI chooser launches the WEKA’s graphical environment which has six buttons: Simple CLI, Explorer, Experimenter, Knowledge Flow, ARFFViewer, & Log.

5.1 Preprocessor

The Preprocess panel imports the data from a database, a CSV file, ARFF etc., and preprocesses this data using filtering algorithm which can be used to transform the data from one format to other e.g. numeric attributes into discrete ones. It is also possible to delete instances and attributes according to specific criteria on the preprocess screen. It is also possible to view the graph for particular attribute.

5.2 CLUSTERING

The Cluster panel is used to access the clustering techniques in Weka, e.g., the simple kmeans, Hierachical, EM, DBScan, cobweb,farthest filter, OPTICS, XMeans algorithm. Sometimes it is necessary to ignore some attribute while using the clustering algorithm, so it is possible with Ignore Attribute button. The proposed work experimented using DBScan clustering.
Figure 4. K-means clustering algorithm result

Figure 5. Complete linkage using Hierarchical clustering.

Figure 6. Result of DBscan is shown in the graph.

Figure 7. Result of cobweb hierarchical tree graph.

Figure 8. Result of EM Algorithm.

Figure 9. Filter cluster algorithm.

Figure 10. Result of optics.

Figure 11. Final work of performance.
5.3 HOT SPOT TECHNIQUE

The Associate panel gives access to association rule e.g. Apriori, Predictive, Apriori algorithm. Once the appropriate parameter for association rule is chosen then result list allows the result set to viewed or saved. Hot Spot learns a set of rules (displayed in a tree-like structure) that maximize/minimize target variables. With a nominal target, one might want to look for segments of the data where there is a high probability of a minority value occurring (given the constraint of a minimum support). For a numeric target, one might be interested in finding segments where this is higher on average than in the whole data set. From the input dataset, the association to the target attributes was found increased. Figure 12 depicts the association with the input and the target attribute. We fix the normal branching factor as 2. It shows that there is some improvement in behavior, attitude and also in student faculty relationship. In BRICHcluster, the maximum cluster is generated by the given input dataset is 9 from X0 to X9. It is relatively optimal, and it is predicted that ratio of the dataset is found strengthened. This will help to analyze the noisy factor, which is low and can be eliminated by adopting the tutoring system in higher education combining of all the year students to a tutor.

VI. CONCLUSION

In this dissertation make use of data mining process in a students’ database using different clustering methods to predict Urban and Rural students’ learning activities. The research information generated after the implementation of data mining technique may be helpful for instructor as well as for students. This work may improve student’s performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of education. In the recent few years data mining techniques covers every area in our life.

Data mining techniques in mainly in the medical, banking, insurances, education etc. before start working in the with the data mining models, it is very necessary to knowledge of available algorithms. The research work explored prediction and analysis of Urban and Rural student improvement, behavior, attitude and relationship of student and staff members. IBARC Algorithm discussed about the prediction and elimination of cluster patterns which is not similar to optimal attribute relations. Weka tool preprocessed the student dataset which showed percentile of the student of higher education had good behavior, average attitude and optimal relationship with staff and tutor. The maximum clusters found to be generated using BRICH Clusters and DBScan analysis. It exhibits the system to the improvement of teaching learning process towards the academic improvement of students in higher education.

In this paper, clustering technique is used for finding the similarity between urban and rural students programming skills. K-means clustering is used to determine the programming skills of the students. This study clearly indicates that the rural and the urban students differ in their programming skills. The huge proportions of urban students are good in programming skill compared to rural students. It divulges that academicians provide extra training to urban students in the programming subject.
REFERENCES


