Data Mining in Educational System using WEKA

Sunita B Aher
ME (CSE) Student
Walchand Institute of Technology, Solapur

Mr. LOBO L.M.R.J.
Associate Professor & Head, Department of IT
Walchand Institute of Technology Solapur

ABSTRACT
Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential used in various commercial applications including retail sales, e-commerce, remote sensing, bioinformatics etc. Education is an essential element for the progress of country. Mining in educational environment is called Educational Data Mining. Educational data mining is concerned with developing new methods to discover knowledge from educational database. In order to analyze student trends & behavior towards education an attempt to study the present behavioral pattern of student in a cross section is a must. This paper surveys an application of data mining in education system & also present result analysis using WEKA tool. As we know large amount of data is stored in educational database, so in order to get required data & to find the hidden relationship, different data mining techniques are developed & used. There are varieties of popular data mining task within the educational data mining e.g. classification, clustering, outlier detection, association rule, prediction etc.How each of data mining tasks can be applied to education system is explained. In this paper we analyze the performance of final year UG Information Technology course students of our college & present the result which we have achieved using WEKA tool.

Keywords
Classification, Clustering, Association rule, Outlier detection, WEKA.

INTRODUCTION
Now a days, large quantities of data is being accumulated. Seeking knowledge from massive data is one of the most desired attributes of Data Mining. Data could be large in two senses: in terms of size & in terms of dimensionality. Also there is a huge gap from the stored data to the knowledge that could be construed from the data. Manual data analysis has been around for some time now, but it creates a bottleneck for large data analysis. The transition won't occur automatically; in this case, we need the data mining. Data Mining could help in a more in-depth knowledge about the data.

In Credit ratings/targeted marketing, given a database of 100,000 names, we could answer which persons are the least likely to default on their credit cards & identify likely responders to sales promotions. In Fraud detection, we would deal with which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer. In Customer relationship management, we would answer which of customers are likely to be the most loyal, and which are most likely to leave for a competitor. Data Mining is a non-trivial process of identifying valid, novel, useful and ultimately understandable patterns in data. Alternative names for data mining are Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc. Data mining can be used in various applications [3]:

Banking: loan/credit card approval, predict good customers based on old customers, view the debt and revenue changes by month, by region, by sector, and by other factors , access statistical information such as maximum, minimum, total, average, trend, etc.

Telecommunication industry: identify potentially fraudulent users and their atypical usage patterns, detect attempts to gain fraudulent entry to customer accounts, discover unusual patterns which may need special attention, find usage patterns for a set
of communication services by customer group, by month, etc., promote the sales of specific services, improve the availability of particular services in a region.

Retail Industry: Identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies

DNA analysis: compare the frequently occurring patterns of each class (e.g., diseased and healthy), identify gene sequence patterns that play roles in various diseases

Data mining models & tasks are shown in figure 1:

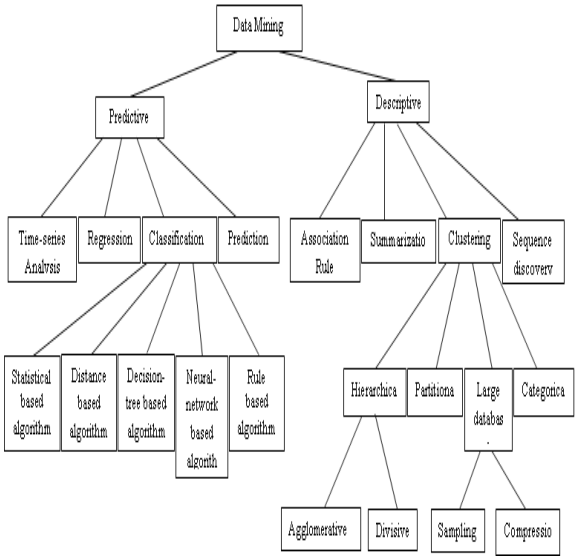

Figure 1: Data mining model & task

2. DATA MINING
Data mining is the process of discovering interesting knowledge from large amount of data stored in database, data warehouse or other information repositories. Based on this view, the architecture of a typical system has the following major components [3] as shown in figure 2:

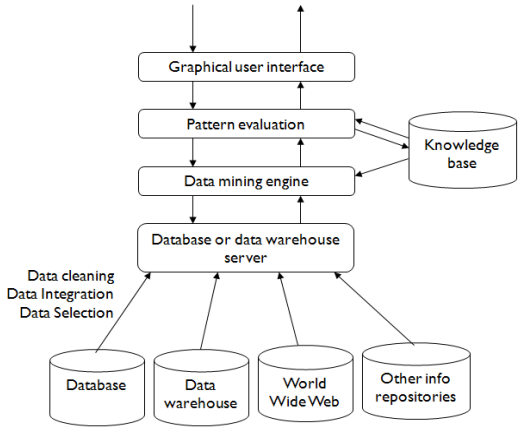

Figure 2: Architecture of a typical Data Mining System [4]

Database, data warehouse, World Wide Web, or other information repository is one or the set of the databases, data warehouse, spreadsheets, or other kinds of information repositories. Data cleaning & data integration techniques may be performed on the data. Database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request. Knowledge base is the domain knowledge that is used to guide the search or evaluate the interestingness of the resulting pattern. Such knowledge can include the concept hierarchy & user beliefs. Data mining engine is essential to the data mining system & ideally consist of set of functional module for tasks such as characterization, association & correlation analysis, classification, prediction, cluster analysis, outlier analysis & evolution analysis. Pattern evaluation module is a component that typically includes interestingness measures & interacts with the data mining modules so as to focus the search towards interesting pattern. The pattern evaluation method can be integrated with data mining module depending on the implementation method used. User interface communicate between the user & the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search & performing the exploratory data mining based on the intermediate data mining results.

Many people treat the data mining as a synonym for another popularly used term, Knowledge Discovery from Data. Others view the data mining as simply an essential step in the process of knowledge discovers [4] as shown in figure 3.

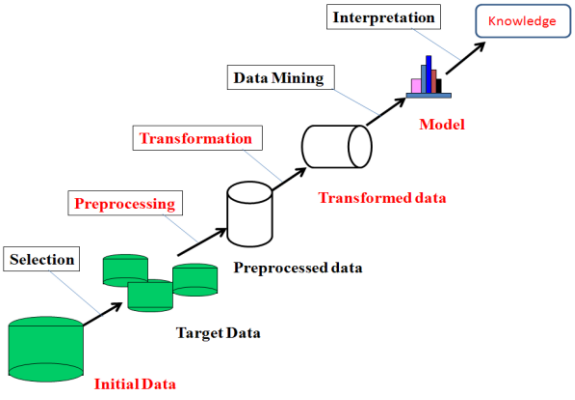

Figure 3: KDD process

The KDD process includes selecting the data needed for data mining process & may be obtained from many different & heterogeneous data sources. Preprocessing includes finding incorrect or missing data. There may be many different activities performed at this time. Erroneous data may be corrected or removed, whereas missing data must be supplied. Preprocessing also include: removal of noise or outliers, collecting necessary information to model or account for noise, accounting for time sequence information and known changes. Transformation is converting the data into a common format for processing. Some data may be encoded or transformed into more usable format. Data reduction, dimensionality reduction (e.g. feature selection i.e. attribute subset selection, heuristic method etc) & data transformation method (e.g. sampling, aggregation, generalization etc) may be used to reduce the number of possible data values being considered. Data Mining is the task being performed, to generate the desired result. Interpretation/Evaluation is how the data mining results are presented to the users which are extremely important because the usefulness of the result is dependent on it. Various visualization & GUI strategies are used at this step. Different kinds of knowledge requires different kinds of representation e.g. classification, clustering, association rule etc

Data mining function & products includes database, information retrieval, statistic, algorithm, & machine learning as shown in figure 4:

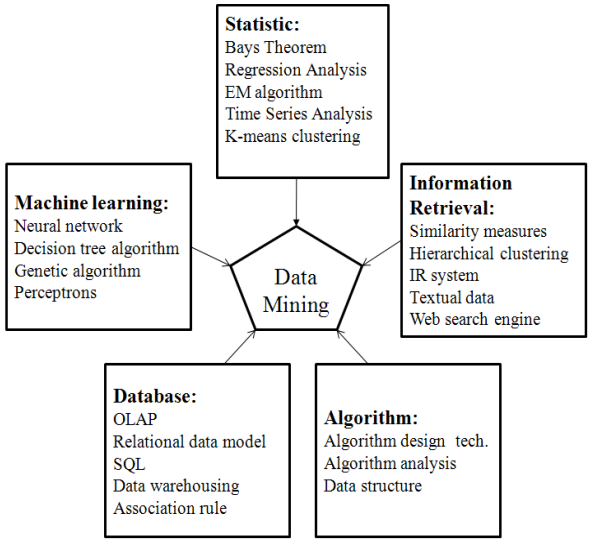

Figure 4: Data mining function & products

3. EDUCATIONAL DATA MINING

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 [8]. Data mining is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. As we know large amount of data is stored in educational database, so in order to get required data & to find the hidden relationship, different data mining techniques are developed & used. There are varieties of popular data mining task within the educational data mining e.g. classification, clustering, outlier detection, association rule, prediction etc. We can use the data mining in educational system as: predicting drop-out student,

relationship between the student university entrance examination results & their success, predicting student's academic performance, discovery of strongly related subjects in the undergraduate syllabi, knowledge discovery on academic achievement, classification of students' performance in computer programming course according to learning style, investing the similarity & difference between schools.

3.1 Association Rule

Association rules are used to show the relationship between data items. Mining association rules allows finding rules of the form: *If antecedent then (likely) consequent* where *antecedent* and *consequent* are itemsets which are sets of one or more items. Association rule generation is usually split up into two separate steps: First, minimum support is applied to find all frequent itemsets in a database. Second, these frequent itemsets and the minimum confidence constraint are used to form rules. Figure 5 shows the generation of itemsets & frequent itemsets where the minimum support count is 2

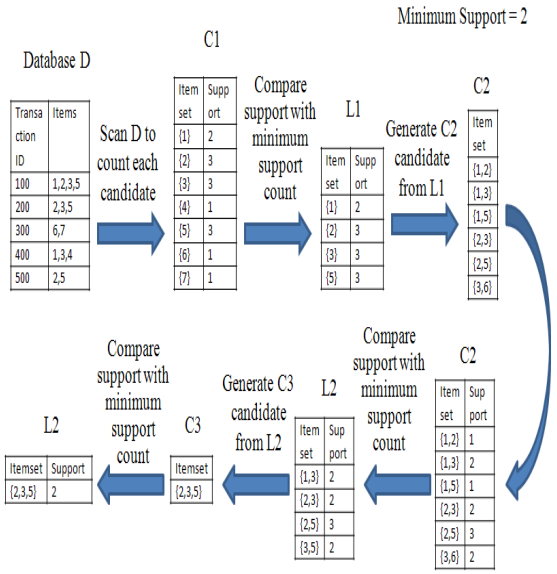

Figure 5: Generation of itemsets & frequent itemsets

Support & confidence are the normal method used to measure the quality of association rule. Support for the association rule X->Y is the percentage of transaction in the database that contains XUY. Confidence for the association rule is X->Y is the ratio of the number of transaction that contains XUY to the number of transaction that contain X.

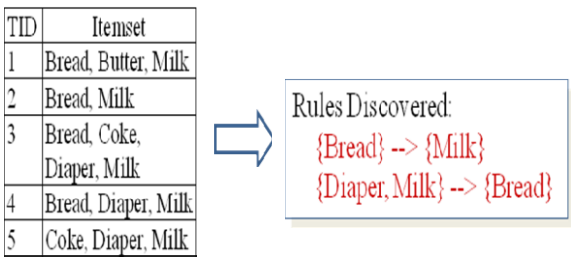

Figure 6: Association rule

Association rule can be used in educational data mining for analyzing the learning data.

3.2 Classification

Classification is a data mining task that maps the data into predefined groups & classes. It is also called as supervised learning .It consists of two steps:

1. Model construction: It consists of set of predetermined classes. Each tuple /sample is assumed to belong to a predefined class. The set of tuple used for model construction is training set. The model is represented as classification rules, decision trees, or mathematical formulae. This model is shown in figure 7.
2. Model usage: This model is used for classifying future or unknown objects. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur. This model is shown in figure 8.

In educational data mining, given works of a student, one may predicate his/her final grade .The decision tree is used to represent logical rules of student final grade

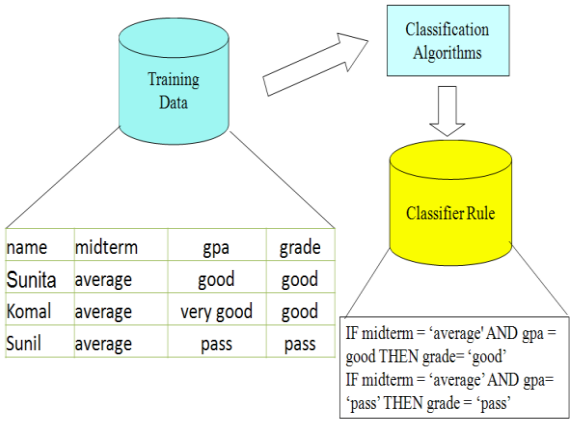

Figure 7: Learning step or model construction

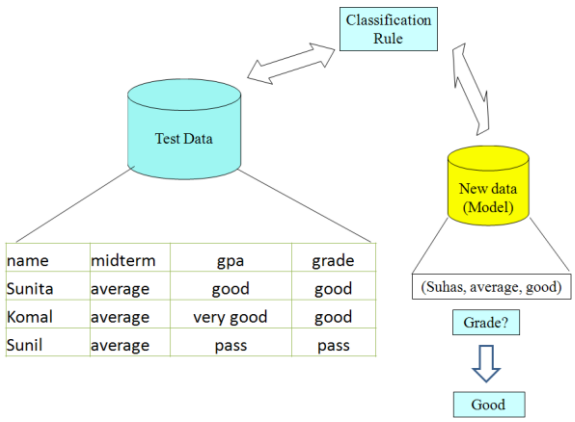

Figure 8: Model Usage (Classification)

3.3 Prediction

It is used to model continuous-valued functions, i.e., predicts unknown or missing values. In this model we deduce single aspect of data from some combination of other aspect of data. In educational data mining prediction can be used to detect student behavior, predicting or understanding student educational outcomes. This model is shown in figure 9.

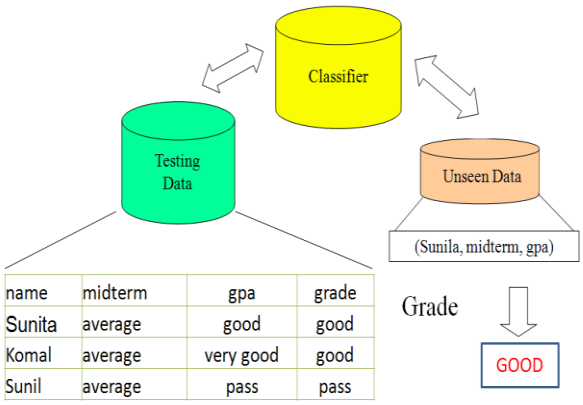

Figure 9: Prediction model

3.4 Clustering

Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group [5] Clustering can be considered the most important unsupervised learning technique. Clustering & its classification is shown in figure 10.

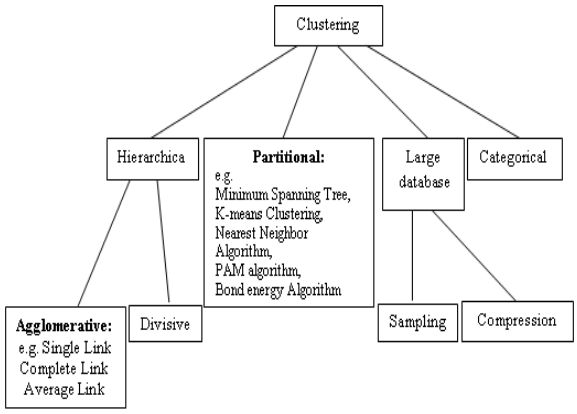

Figure 10: Classification of clustering algorithm

In educational data mining, clustering has been used to group the students according to their behavior e.g. clustering can be used to distinguish active student from non-active student according to their performance in activities.

4. RELATED WORK

The paper [1] presents an approach to classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system. They design, implement, and evaluate a series of pattern classifiers and compare their performance on an online course dataset. Four classifiers were used to segregate the students. The combination of multiple classifiers leads to a significant improvement in classification performance. They used the genetic algorithm (GA) to improve the prediction accuracy & using the genetic algorithm the accuracy of combine classifier performance is about 10 to 12% as compared to the non-GA. This method is of considerable usefulness in identifying students at risk early, especially in very large classes, and allow the instructor to provide appropriate advising in a timely manner. In paper [2], they analyzed how association rule are useful in Educational data mining for analyzing learning data. They explained the cosine & added value (or equivalently lift) are well suited to educational data & that teacher can interpret their result easily. They provide the case study with data from LMS (Learning Management System). In paper [5] they explained how data mining is useful in higher education particularly to improve the performance of the student. For that they used the Database course & also collected all available data including their usage of Moodle e-learning facility. They used association rule, classification rule using decision tree, clustered the student into the group using EM-clustering & using outlier analysis detected the outlier in the data. They used this knowledge to improve the performance. In paper [6], they studied the relationship between the student university entrance examination result & their success using cluster analysis & K-means algorithm techniques. They were grouped the university student according to their characteristic, forming cluster & clustering process carried out using the K-means clustering. In paper [7], they surveyed the application of data mining to traditional educational system particularly web-based courses, intelligent web-based educational system, learning content management system. Each of these system used data source & objectives for knowledge discovery. In each case data mining technique such as statistics & visualization, clustering, classification, outlier detection, association rule mining pattern mining & text mining were applied. They explained how the cycle of applying data mining in educational system as shown

below worked. The cycle of applying data mining in educational system is shown in figure 11.

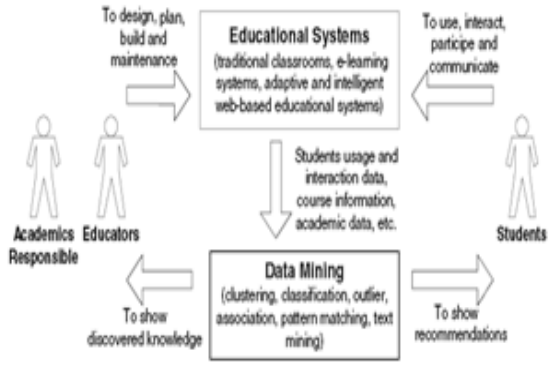

Figure 11: The cycle of applying data mining in educational system [7]

In the [9] they explained how associative classification & clustering is effective in finding the relation & association between the students. They evaluated the student progress according to the association between different factors using the data collected. In paper [10] they were analyzed the log file of elementary school student studied with science web-based module. They also produced Learnogram-the graphical representation tool that visualized the student learning process for each student.

5. WEKA TOOL

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 [11]. 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.

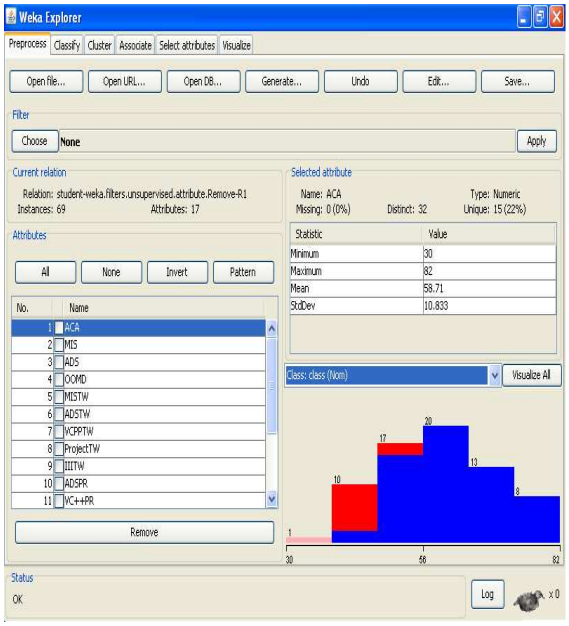

Figure 12: Weka 3.5.3 with Explorer window open with Student dataset

The *Explorer* interface has several panels that give access to the main components of the workbench:
1. 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.
2.The *Classify* panel allows the user to apply classification and regression algorithms (e.g. NaiveBays algorithm, ADTree, ID3 Tree , J48 Tree, ZeroR rules etc) to the dataset estimate the accuracy of the resulting model**.** It is also possible to visualize erroneous predictions, ROC curves, etc. Result of classification can be seen in classifier output area
3. The *Cluster* panel is used to access the clustering techniques in Weka, e.g., the simple k-means**,** EM, DBScan, XMeans algorithm. Sometimes it is necessary to ignore some attribute while using the clustering algorithm, so it is possible with Ignore Attribute button.
4. The *Associate* panel gives access to association rule e.g. Apriori, PredictiveApriori algorithm. Once the appropriate parameter for association rule is chosen then result list allows the result set to viewed or saved.
5. The *Select attributes* panel allows to search among all possible combination of attribute in dataset, which subset of attribute is best for making prediction.
6. The *Visualize* panel visualizes 2D plots of current relation.

6. EXPERIMENTAL WORK

We consider the result of final year of UG Information Technology of our institute for result analysis. There are four Subjects namely ACA (Advanced Computer Architecture), MIS (Management Information System), ADS (Advanced Database System), OOMD (Object Oriented Modeling & Design).The graph for each subject, total, percentage, class, pass subject & fail subject is shown in figure 12. As in figure 12, consider the graph for class attribute, 58 students got the distinction, 10 students-the first class & 1 student-ATKT. In the same way, each graph can be analyzed.

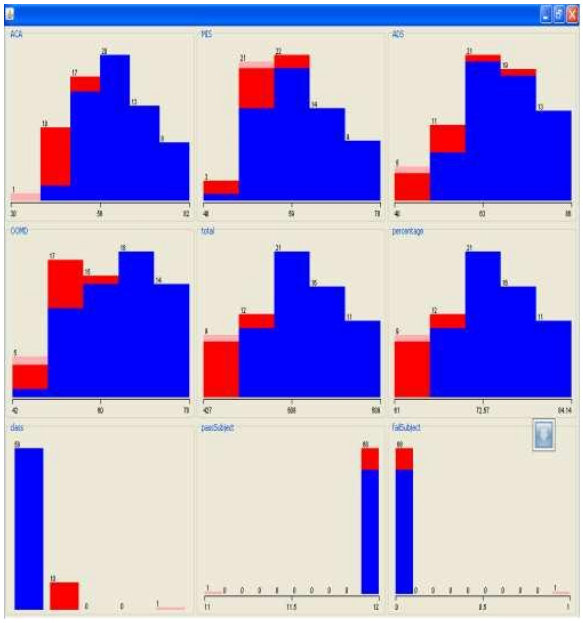

Figure 13: Result from Weka

In WEKA, ZeroR classifier predicts the majority of class in training data. It predicts the mean for numeric value & mode for nominal class. As shown in figure 13, the confusion matrix is given which gives the accuracy of solution to the classification problem. As can be seen from the figure 13, 58 students got the first class, 10-first class & 1-ATKT.

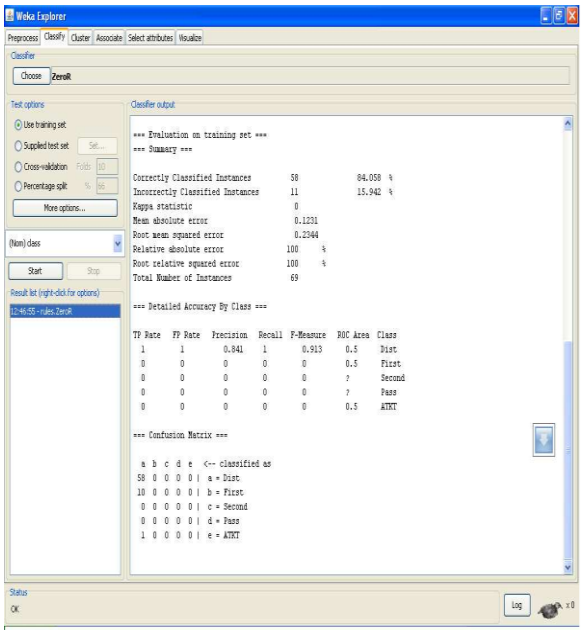

Figure 14: Result of Classifier

Now consider the clustering algorithm DBSCAN(Density Based spatial clustering of applications with noise). This algorithm is used to create the cluster with a minimum size & density. This algorithm also handles the outlier problem.The result of training set using this algorithm is given in figure 14 .

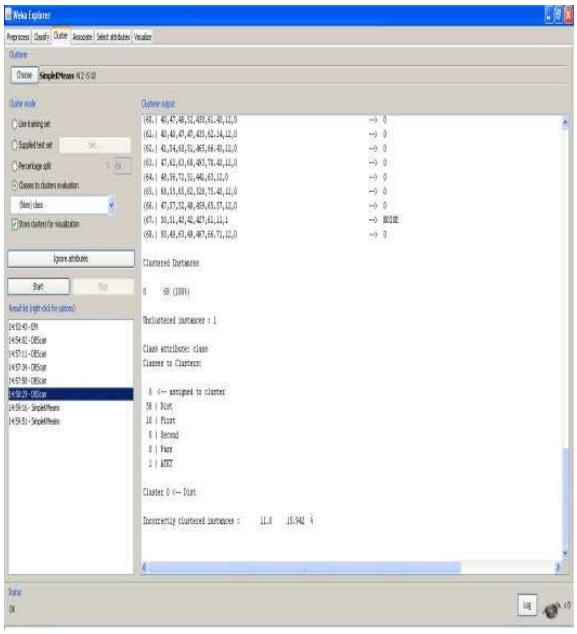

Figure 15: Result of Clustering algorithm DBSCAN

7. CONCLUSION & FUTURE WORK

In this paper, we have studied how data mining can be applied to educational systems. It shows how useful data mining can be in higher education, particularly to improve students' performance. We used students' data from the database of final year students' for Information Technology UG course. We collected all available data including their performance at university examination in various subjects. We applied data mining techniques to discover knowledge. We discovered classification using ZeroR algorithm. Also we clustered the student into group using DBSCAN-clustering algorithm. Finally, noisy data was detected. Each one of this knowledge can be used to improve the performance of student.

For future work, a way to generalize the study to more diverse courses to get more accurate results needs to be developed. Also, experiments could be done using more data mining techniques such as neural nets, genetic algorithms, k-nearest Neighbor, Naive Bayes, support vector machines and others. Finally, the used preprocessed and data mining algorithms could be embedded into e-learning system so that one using the system can be benefited from the data mining techniques.

8. REFERENCES

[1] Behrouz.et.al., (2003) Predicting Student Performance: An Application of Data Mining Methods with The Educational Web-Based System Lon-CAPA © 2003 IEEE, Boulder, CO

[2] Sheikh, L Tanveer B. and Hamdani, S., "Interesting Measures for Mining Association Rules". IEEE-INMIC Conference December. 2004

[3] Han,J. and Kamber, M., "Data Mining: Concepts andTechniques", 2nd edition.

[4] *"Data Mining* Introductory and Advanced Topics" byMargaret H. *Dunham*

[5] Alaa el-Halees (2009) Mining Students Data to Analyze e-Learning Behavior: A Case Study.

[6] Erdogan and Timor (2005) A data mining application in a student database. Journal of Aeronautic and Space Technologies July 2005 Volume 2 Number 2 (53-57)

[7] Romero C. and Ventura S.,"Educational data mining: A Survey from 1995 to 2005".Expert Systems with Applications (33) 135-146. 2007

[8] International Educational Data Mining Societywww.educationaldatamining.org

[9] Kifaya (2009) Mining student evaluation using associative classification and clustering Communications of the IBIMA vol. 11 IISN 1943-7765

[10] Galit.et.al (2007) Examining online learning processes based on log files analysis: a case study. Research, Reflection and Innovations in Integrating ICT in Education

Sunita B Aher received the B.E degree in Computer Science & Engineering in 2000 from Amravati University, Amravati, India and pursuing the M. E. degree in Computer Science in Walchand Institute of Technology, Solapur, India.

She is doing the dissertation work under the guidance of Mr. Lobo l.M.R.J, Associate Professor & Head, Department of IT, Walchand Institute of Technology, Solapur, Maharashtra, India.

Mr. Lobo l.M.R.J received the B.E degree in Computer Engineering in 1989 from Shivaji University, Kolhapur, India and the M. Tech degree in Computer and Information Technology in 1997 from IIT, Kharagpur, India.

He has registered for Ph.D in Computer Science and Engineering at SGGS, Nanded of Sant Ramanand Teerth Marathawada University, and Nanded, India. Under the guidance of Dr. R.S. Bichkar. He is presently working as an Associate Professor & Head, Department of IT Walchand Institute of Technology, Solapur, Maharashtra, India. His research interests include Evolutionary Computation, Genetic Algorithms and Data Mining.