Comparative Analysis of Data Mining Techniques on Educational Dataset

Sumit Garg
M.Tech Scholar
Dept. of Computer Science
Shekhawati Engineering College
Dundlod, Rajasthan, India

Arvind K. Sharma
Guest Faculty
Dept. of Computer Science
University of Kota, Kota
Rajasthan, India

ABSTRACT

Data mining is a relatively young and interdisciplinary field of computer science. It is a process that attempts to discover new patterns in large data sets. Different types of mining algorithms have been proposed by different researchers in recent years. A single algorithm may not be applied to all applications due to difficulty for suitable data types of the algorithm. Therefore the selection of a correct data mining algorithm depends on not only the goal of an application, but also on the compatibility of the data set. The aim of this paper is how to use suitable data mining algorithms on educational dataset. This paper focuses on comparative analysis of various data mining techniques and algorithms.

General Terms

Data Mining

Keywords

Data Mining Techniques, Educational Dataset, WEKA

1. INTRODUCTION

The field of data mining is an emerging research area with important applications in Engineering, Science, Medicine, Business and Education. The size of data base in educational application is large where the number of records in a data set can vary from some thousand to thousand of millions. The size of data is accumulated from different fields exponentially increasing. Data mining has been used different methods at the intersection of Machine Learning, Artificial Intelligence, Statistics and Database Systems[1]. The overall aim of the data mining process is to extract information from huge datasets and transform it into understandable structure for further use. Data mining techniques which extract information from huge amount of data have been becoming popular in education domains.

The rest of paper is organized as follows:

Section 2 describes literature review in brief. Section 3 explains various types of data mining techniques. Section 4 contains proposed methodology. Section 5 summarizes the comparative analysis of different data mining techniques and algorithms. Conclusion is shown in section 6 while references are mentioned in the last section.

2. LITERATURE REVIEW

This section summarizes various review and technical articles on data mining techniques applied in education field. Several works have been carried out by many researchers. This section presents a brief summary on the basis of literature.

In [2] Velmurgun T. et al. attempted to analyze performance of K-means and Fuzzy C-means clustering techniques in the field of data mining. The performance compared on the basis of clustering result quality.

In[3] Kavitha P., T. Sasipraba evaluated the performance of distributed data mining framework on Java platform. Association rule mining was used for discovering interesting patterns from a large amount of data.

In[4] Yujie Zheng proposed a methodology for clustering in data mining to improve the standard of higher education used to find data segmentation and pattern information.

In[5] M.Sukanya et al. used classification and clustering algorithms of data mining for the performance improvement in education sector. By using these algorithms an educational institute could predict the number of enrolled students.

In[6] Manoj Bala et al. applied an application of data mining in educational institute to extract the useful information from the huge dataset and provided analytical tool to view and used this information for decision making process. They also conducted a research on student learning result based on data mining.

In[7] Hamidah Jantan used a potential classification technique for academic talent forecasting in higher educational institutes. He proposed a classification model to increase the academic talent in higher educational institutions.

In[8] Tai Chang Hsia applied data mining techniques to analyze the course preferences and course completion rates of enrollees in extension education courses at a University in Taiwan. Some of the data mining algorithms like decision tree, link analysis, and decision forest were used for further analysis.

3. DATA MINING TECHNIQUES

Generally data mining contains several algorithms and techniques for picking out interesting patterns from large data sets. Data mining techniques are classified into two categories: supervised learning and unsupervised learning.

In supervised learning, a model is built prior to the analysis. We then apply the algorithm to the data in order to estimate the parameters of the model. Classification, Decision Tree, Bayesian Classification, Neural Networks, Association Rule Mining etc. are common examples of supervised learning.

In unsupervised learning, we do not create a model or hypothesis prior to the analysis[9]. We just apply the algorithm directly to the dataset and observe the results. Then a model can be created on the basis of the obtained results. Clustering is one of the examples of unsupervised learning. Various data mining techniques such as Classification, Decision Tree, Bayesian Classification, Neural Networks, Clustering, Association Rule Mining, Prediction, Time Series Analysis, Sequential Pattern and Genetic Algorithm and Nearest Neighbour have been used for knowledge discovery from large data sets[10]. Some of the common and useful data mining techniques have been discussed.

3.1 Classification

Classification is a supervised learning technique. It maps the data into predefined groups. It is used to develop a model that can classify the population of records at large level. Classification algorithm requires that the classes be defined based on the data attribute value. It describes these classes according to the characteristics of the data that is already known to belong to the classes. The classifier training algorithm uses these pre-defined examples to determine the set of parameters required for proper discrimination. This section discusses some of the useful data mining techniques such as Decision Tree, Neural Networks, Bayesian Classification etc.

3.1.1 Decision Tree

A decision tree is a flow chart like tree structure, where each node denotes test on an attribute value, each branch represents the result of the test, and tree leaves represent classes. The drive model can be represented in different forms such as classification (If-Then) rules, decision tree, mathematical formula or neural networks. Decision tree can easily be converted to classification tree[10]. Decision trees are simple to understand and provide good results even with small data. Decision tree induction algorithms can be used for classification in many application areas, such as Education, Medicine, Manufacturing, Production, Financial analysis, Fraud Detection and Astronomy etc. There are several data mining algorithms such as C4.5, ID3, CART, J48, NB Tree, REP Tree etc.

3.1.2 Bayesian Classification

Bayesian classifier is statistical classifier. It can be used to predict class membership probabilities. This probability about the tuple that belongs to the particular class or not. Bayesian classification is based on Bayes theorem. Suppose X is a data tuple. It is considered as 'evidence'. H is the hypothesis that the data tuple X belongs to a specified class C. We also determine $P(H/X)$ the probability that the hypothesis H holds given the evidence or observe data tuple X[11]. $P(H/X)$ is the posterior probability or a posteriori probability of H conditioned on X. For example, the data tuple is confined to the University described by attribute placement and results and X is a University with placements are good and results are average. Suppose H is a hypothesis that the university gets A grade from UGC. Then $P(H/X)$ is the probability that University X will get A grade and results and placement are known. In contrast $P(H)$ is the prior probability of H. For example, this probability that any given University will get A grade, regardless of placement and results. The posterior probability $P(H/X)$ is based on more information about the University in comparison of prior probability $P(H)$, which is independent of X. Similarly, P(X/H) is the posterior probability of X conditioned on H. That is University has good placement and average results, and University will get A grade. $P(H)$, $P(X/H)$ and $P(X)$ may also be estimated from given data. Bayes theorem provides a useful way of calculating the posterior probability, $P(H/X)$ from $P(H)$, $P(X/H)$ and $P(X)$. Bayes theorem is as follows:

$$P\,(H/X) = \frac{P(X/H)\;P\,(H)}{P(X)}$$

3.1.3 Neural Networks

The area of neural networks probably belongs to the border line between the artificial intelligence and approximation algorithm. A neural network is a collection of neurons like processing units with weighted connection between the units. It composes of many elements, called nodes which are connected in between. The connection between two nodes is weighted and by the adjustment of this weight, the training of the network is performed[12]. A classification model can be represented in different forms like neural network and decision tree which is shown in fig. 1(a) and fig.1(b). There are many advantages of neural networks such as adaptive learning ability, self-organization, real time operation and insensitivity to noise. Neural networks are used to identifying patterns or trends in data and well suited for prediction or forecasting needs. There are several neural network algorithms such as Back Propagation, NN Supervised Learning, and Radial Base Function (RBF) Network etc.

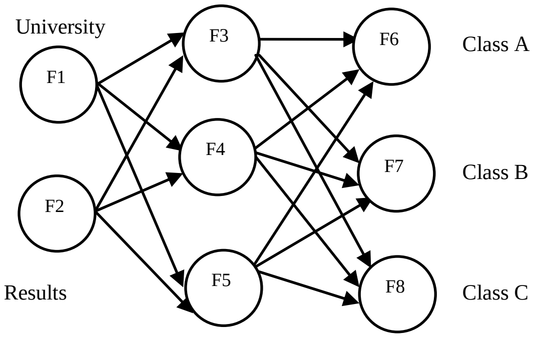


Fig.1(a) : Neural Network

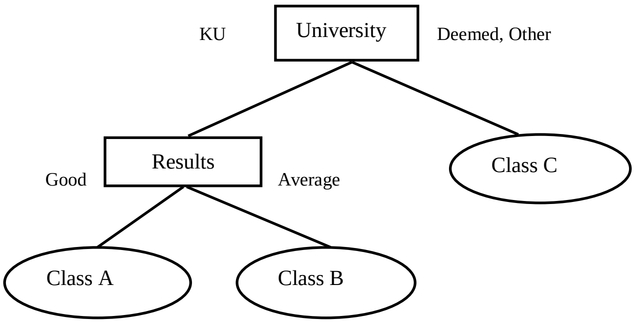


Fig.1(b) : Decision Tree

3.2 Clustering

Clustering as the name suggests is the process of grouping data into classes, so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other cluster. Dissimilarities have been observed on the basis of attribute value describing the objects often distance used. It is the process of identification of similar classes of object. It has been used as a grouping a set of physical or abstract objects into classes of similar objects. A cluster is a collection of data objects which are similar to another within the same cluster and dissimilar to the object in other cluster. It is an unsupervised learning technique that shows the natural groupings in data. Clustering has been frequently used in data mining applications for discovering patterns in huge datasets[9]. There are many clustering techniques like Partitioning methods(K-means, K-medoids), Hierarchical methods(CURE, CHAMELEON), Density based methods(DBSCAN & OPTICS), Grid based methods (STING, CLIQUE) and Model based methods(EM algorithm).

3.3 Association Rule Mining

Association rule mining is the discovery of association relationships or correlation among a set of items. Association and correlation is used to find the frequent item set among large data sets[11]. Association rule for a given dataset is very large and they are generally in (if any) value. The main task of association rule mining is to find sets of binary variable that co-occur together frequently in a transaction database[12]. Association rule holds many algorithms like Apriori, CDA, DDA, interestingness measure. Association rules are *if then* statements that search uncover relationship between unrelated data in the relational database. Many association rule mining techniques exist on the basis of literature such as multilevel association rule, Multi dimensional association rule and quantitative association rules.

3.4 Prediction

Prediction is a data mining technique that is used to identify the relationship between independent variables and relationship between dependent and independent variables[13]. Prediction analysis can be used in education domains. Regression technique can be used to generate a model for prediction. Regression analysis can be used to model the relationship between one or more independent variable and dependent variables[14]. Prediction techniques can be used to predict the possible values of some missing data and the value distribution of certain attributes in a set of objects. It finds the attribute related to the interest and predicting the value distribution based on the set of data similar to the selected objects. There are many regression techniques such as Linear Regression, Nonlinear Regression, Multivariate Linear Regression, and Multivariate Nonlinear Regression.

3.5 Time Series Analysis

It is a sequence of data points, measured typically at successive times spaced at uniform time intervals. Typical examples include stock prices, currency exchange rates, and the volume of product sales, biomedical measurements, and education data etc, collected over monotonically increasing time. Rule induction algorithm such as Version Space, AQ15, C4.5 rules are presently employed in time series data mining applications[15].

3.6 Sequential Patterns

It is one of the data mining techniques that seek to discover similar patterns in data transaction over a business period[13].The uncover patterns are used for further business analysis to recognize relationships among data.

4. PROPOSED METHODOLOGY

The proposed methodology will be used to generate a database for the current study. For this study, we have been collected dataset of B.Tech students for past three years from the Computer Science Department and Examination Cell of an Engineering College, Bharatpur, Rajasthan, India. Before processing of data we will be going to clean the data to remove noise and inconsistency. To remove missing values in the dataset, we will use the cleaning techniques. The experiments and observations will be carried out by using data mining tool i.e. WEKA.

4.1 Data Selection

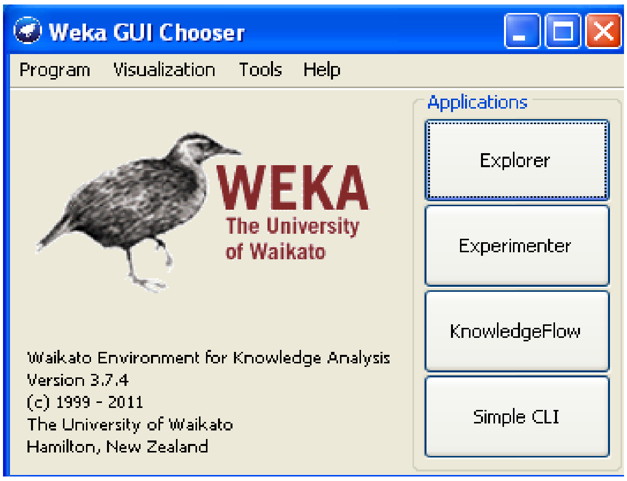
In this paper dataset has been collected from an Engineering College**,** Bharatpur, India. The dataset contains different attributes and instances. The complete description of dataset is shown in table 1.

Table 1: Attributes of Dataset

Attribute	**Description**
Enroll No.	University Enrollment No. of the student
Roll No.	University Roll No. of the student
Final Result	1–Pass, 0–Fail
Sem	Semester of which result is declared
Branch	Branch of the student
Sub	Name of particular subject
Tot ThMarks	Total Theory marks obtained by the student
TotPrMarks	Total Practical Marks Obtained by the student

4.2 Tool Selection –WEKA

WEKA stands for Waikato Environment for Knowledge Analysis. WEKA is a data mining tool developed at the University of Waikato, Newzeland. It uses GNU general public licenses and is freely available on following link: http://www.cs.waikato.ac.nz/~ml/weka[16]. It is implemented in Java programming language and has GUI for loading data, running analysis and producing visualization of result. WEKA supports many data mining techniques and algorithms like classification, clustering, feature selection, data preprocessing, regression, visualization and clustering. The GUI Interface of WEKA is shown in fig. 2.



Waikato Environment for Knowledge Analysis
Version 3.7.4
(c) 1999 - 2011
The University of Waikato
Hamilton, New Zealand

Fig. 2: GUI Interface of WEKA

This is WEKA GUI Chooser. It provides four interfaces to work on:

4.2.1 Explorer

It is used for exploring the data with WEKA by providing access to all the facilities by the use of menus and forms.

4.2.2 Experimenter

WEKA Experimenter allows you to create, analyze, modify and run large scale experiments.

4.2.3 Knowledge Flow

It has the same function as that of explorer. It supports incremental learning. It handles data on incremental basis. It uses incremental algorithms to process data.

4.2.4 Simple CLI

CLI stands for command line interface. It just provides all the functionality through command line interface.

5. COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES & ALGORITHMS

This section presents the comparative analysis of different data mining techniques and algorithms which have been used by most of the researchers in educational data mining. A brief summary of these data mining algorithms with their merits and demerits have been discussed.

The comparative study of classification algorithms such as Decision Tree, Naïve Baysian and Neural Networks is shown in table 2.

Table 2: Comparison of Classification Algorithms

Algorithm	Merits	Demerits
Decision Tree	• It can handle both continuous and discrete data. • It provides fast result in classifying unknown records. • It works well with redundant attribute. • It provides good results with small size tree. Results does not affect with outliers. • It does not require preparation method like normalization. • It also works well with numeric data.	• It can’t predict the value of a continuous class attribute. • It provides error prone results when too many classes are used. • Irrelevant attribute affect construction of decision tree in a bad manner. • Small change in data can change the decision tree completely.
Naïve Bayesian	• It provides high accuracy and speed on large database. • It has minimum error rate in comparison to all other classifier. • It is easy to understand. • It is not sensitive to irreverent features. • It handles streaming data well. • It can also handle real and discrete values.	• It assumes independence of features. So it provides less accuracy.
Neural Networks	• NNs have high tolerance of noisy data. • They are well suited for continuous value. • They can classify pattern on which they have not been trained.	• They have poor interpretability. • They contain long training time.

The comparative study of common decision tree algorithms such as C4.5, ID3, J48, CART and J48 is shown in table 3.

Table 3: Comparison of Decision Tree Algorithms

Algorithm	Merits	Demerits
C4.5	• It uses continue data. • It avoids over fitting of data. • It improves computational efficiency. • It handles training data with missing and numeric value.	• It requires that target attribute will have only discrete values.
ID3	• It builds the fastest and short tree. • Understandable prediction rules are created from the training data. • Finding leaf nodes enable test data to be pruned, reducing number of test.	• It can’t handle numeric attributes and missing values. • Data may be over-fitted and over-classified, if a small sample is tested. • Only one attribute at a time is tested for making a decision.
CART	• It is non-parametric. • It does not require variable to be selected in advance. • It easily handles outliers.	• It may have unstable decision tree. • It splits only by one variable.
J48	• It can handle both nominal and numeric values. • It can also handle missing values.	--

The comparative study of common classification and clustering algorithms such as K-NN and K-means is shown in table 4.

Table 4: Comparison of K-NN and K-means Algorithms

Algorithm	Merits	Demerits
K-Nearest Neighbor	• It performs better with missing data • It is easy to implement and debug. • It provides more accurate results. • Some noise reduction techniques are used that improve the accuracy of classifier.	• It has poor run time performance. • It requires high calculation complexity. • It considers no weight difference between samples. • It is sensitive to irrelevant and redundant feature.
K-Means	• It is Reasonable fast. • It is very simple and robust algorithm. It provides best results when data sets are distinct.	• It can't work with non-linear data sets. • It can't handle noisy data and outliers.

6. CONCLUSION

The result of this paper indicates that capabilities of data mining techniques provide effective improving tools for student's performance in education field. A comparative analysis of various data mining techniques is presented in this paper. Many data mining techniques can be implemented on student's data to predict their future performance. This paper shows how useful data mining techniques can be applied in higher education particularly to predict the final performance of the students. It will help the higher educational institutions to decide the individual programs to enhance skills of the students to improve their performance. In future work, application of data mining techniques in education field will be used to develop a model for performance monitoring and evaluation system.

7. REFERENCES

[1] Bharati M. Ramageri, "Data Mining Techniques and Application", Indian Journal of Computer Science and Engineering, Vol. 1 No. 4; pp. 301-305.

[2] Velmurugan T. et al., "Performance Evaluation of K-Means & fuzzy C-means Clustering Algorithm for Statistical Distribution of Input Data Points", European Journal of Scientific Research, Vol. 46, 2010.

[3] Kavitha P., T. Sasipraba, "Performance Evaluation of Algorithms using a Distributed Data Mining Framework based on Association Rule Mining", International Journal on Computer Science & Engineering (IJCSE), 2011.

[4] Yujie Zheng, "Clustering Methods in Data Mining with its Application in Higher Education", International Conference on Education Technology and Computer, Vol. 43, 2012, IACSIT Press, Singapore.

[5] M.Sukan et al., "Data Mining: Performance Improvement in Education Sector using Classification and Clustering Algorithm", International Conference of Computing and Control Engineering (ICCCE) 12-13 April, 2012.

[6] Manoj Bala et al., "Study of Application of Data Mining Technique in Education", International Journal of Research in Science and Technology, Vol. No. 1, Issue No. IV, Jan-March, 2012.

[7] Hamidah Jantan, "Classification and Prediction of Academic Talent using Data Mining Technique", 14 International Conference on Knowledge based and Intelligent Information and Engineering Information pages 491-500.

[8] Tai Chang Hsia, "Course Planning of Extension Education to meet Market Demand by using Data Mining Techniques", Expert System with Applications: An International Journal, Vol.34, Issue-1, Jan 2008.

[9] Jiawei Han and Michelire Kamber, "Data Mining Concept and Technique", Published by Morgan Kaufman, 2006.

[10] Monika Goyal and Rajan Vohra, "Application of Data Mining in Higher Education", International Journal of Computer Science(IJCSI) Issues, Vol. 9, Issue-2, No.1, March 2012; pp-113-120.

[11] Arun K Pujari, "Data mining Technique", Published by Universities Press (I) Pvt. Ltd, Hyderabad, India.

[12] Gajendra Sharma, "Data mining and Data Warehousing and OLAP", Published by S.K. Kataria & Sons, New Delhi, India.

[13] Arvind Sharma et al., "Data Mining Techniques and Their Implementation in Blood Bank Sector-A Review", International Journal of Engineering Research and Application (IJERA), Vol. 2, Issue-4, July-August 2012; pp.1303-1309.

[14] R.K. Somani, "Data Mining & Warehousing", College Book Centre, Chaura Rasta, Jaipur, India.

[15] Venkatadri.M, "A Review on Data Mining from Past to Future", International Journal of Computer Applications (IJCA), Vol. 15, No.7, Feb 2011.

[16] http://www.cs.waikato.ac.nz/~ml/weka/