Application of Data Mining with Teaching Assistant Evolution: A Review

Akhilesh Kumar Shrivas
Department of IT , Dr. C. V. Raman University
Bilaspur, Chhattisgarh, India

Shikha Mishra
Department of IT , Dr. C. V. Raman University
Bilaspur, Chhattisgarh, India

Abstract - Now days, teaching assistant evaluation are very important for every academic sector like primary, middle, high school as well as higher education to improve the education system. Teaching assistant evaluation plays major role to reformation of academic sector and improve the standard of education. This research work focus on the performance of teaching assistant that is analyzed by the different authors. There are various authors have used different data mining techniques as well as other classification techniques to improve the performance of model. This paper also focus on the data mining and its various application that can be applied for teaching assistant evaluation.

Keywords – Classification, Data Mining, Teaching Assistant Evaluation.

I. INTRODUCTION

Today's teaching assistant is major role for (TAs) [1] individuals who assist teachers or professors with instructional responsibilities. TAs includes Graduate Teaching Assistants (GTAs), Undergraduate Teaching Assistants (UTAs), secondary school TAs and elementary school TAs. The responsibility of teaching assistants is regular interaction with students, take proper classes, motivate students and try to make good learning environment for students. The quality of education is one of the important factor in education and various institute try to reform in education system ,hence quality may be improve. Due to large number of students and competitive environment, good teacher assistant plays important role. Due to large number of information are increasing day by day related to teaching criteria ,the data mining plays very important role.

There are various authors have worked in the field of teaching assistant evaluation. E. G. Dragomir et al. (2010)[6] have proposed K-Nearest Neighbor techniques and Support vector machine for Teaching performance evaluation . Experimental results shows that SVM given highest accuracy as 96.67% as best model. P. kaur et al. (2015)[9] have used various classifier such as multilayer Perception ,Navie bayes ,SMO,J48 and REPTree using WEKA open sources tool for analyze student's performance. The highest accuracy as 82% in case of MLP. A. Gupta et al. (2015)[5] have used various classifications techniques like J48 ,Decision Table, Multilayer Perceptron (MLP), Naive bayes and other algorithms for evaluation of teaching assistant .The authors have applied teaching assistant data that is collecting from UCI Repository and they got highest accuracy (41.05%) in case of J48 and Naive bayes classifier. T. Z. Mohammad et al.(2015)[10] have proposed an intelligent educational data mining classification model designed for teaching English for slow learner's students. The model is also called IEDM-SL. The main motive of IEDM_SL model is to identifying learning pattern and improve their performance. J. Yang et al. (2011) [1] have suggested SVM technique for teaching assistant evaluation. They have trained and tested the SVM model with different kernel functions like linear, polynomial, radial basis function. A. A. Balamurugan Subramanian et al.(2010) [7] have presented many classification techniques . The experiment shown that teaching evaluation data set applied on various classification techniques like C4.5, NB,and k-NN with 10-fold cross validation. The proposed Parkinson decision tree with NB classifier given highest accuracy 98.97%. S. Agrawal et al. (2012) [8] have suggested 8 different classifier for classification of data in WEKA environment and compared among them. The

performance of LIBSVM classification as a best classifier and given accuracy is 97.3%. S. Mardikyan et al. (2011)[13] have discussed about the student evaluation that evaluate the teacher performance in higher education. They have suggested two different factor like stepwise regression , decision tree for analyzing teaching performance.

II. Data Mining

Data mining is one of the import domains that is applying in different areas like information security, health care, market analysis and others. In which education sector is one of the important area where we are applying data mining to extract the useful information, hence we can improve the education quality. Data mining or Knowledge Discovery in Database (KDD) plays important role to extract the useful or important information from huge amount of data. Data mining techniques are the long process of research development.

A. Knowledge Discovery in Database Process (KDD)-

Data mining [2] is known to be a part of Knowledge Discovery process in which data is analyzed and summarized from different perspective and converted into useful information .It helps in extracting the hidden and valid data which has the potential of transformed into useful information. It is similar to machine learning process. Figure 1 shows that the KDD process that consist steps like data cleaning, data integration, data selection, data transformation, data mining ,pattern evolution and knowledge presentation. The data cleaning, data integration, data selection, data transformation is called data preprocessing. The next process is called data mining that extract the useful pattern from large database and present using various tools like bar chart, pie chart, line chart etc.

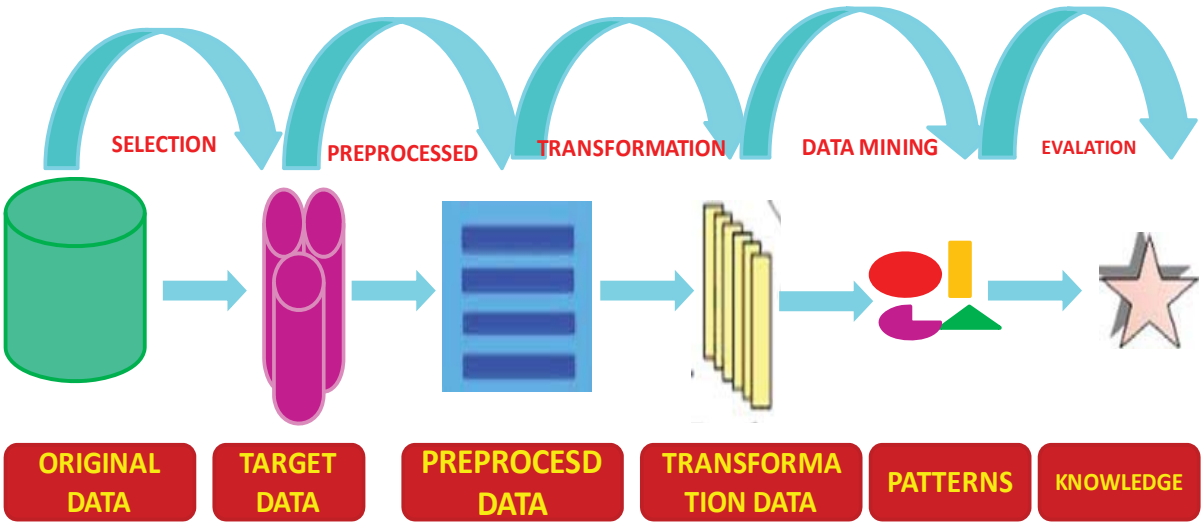

Figure 1: (KDD) Knowledge Discovery in Database Process

III. DATA Mining Applications

Data mining is one of the biggest and important research areas that can be applied in various fields for analysis of data. There are various applications of data mining techniques:

A. Clustering-

Clustering is one of the important data mining applications that is used to form a group as clusters. Clustering is technique to form groups of similar data or objects. The process of grouping a set [2] of physical or abstract objects into classes of similar object is called clustering .A cluster of data objects can be treated collectively as one group in many application. Cluster analysis has been widely used in numerous applications, including pattern recognition

,data analysis ,image processing ,and market research.

B. Association Rule Mining -

Association rule mining (ARM) is a data mining application that is used to find the frequent item set transaction in database. ARM finds the relationship among items in large dataset. Among huge amounts of business transaction record can help in many business decision making processes such as catalog design, cross-marketing and loss leader analysis [2].

C. Prediction-

Classification and prediction [2] are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. The main different between classification and prediction is that classification predicts categorical labels and prediction models continuous valued function.

D. Classification-

Classification is one of the important applications of data mining techniques for classification of data. Classification is also called supervised learning because each record associated with class. Classification process consist two steps: first step is trained the model using training samples of data set and second one is testing the trained model using testing sample of same data set. Figure 2 shows that generic architecture of training and testing the model.

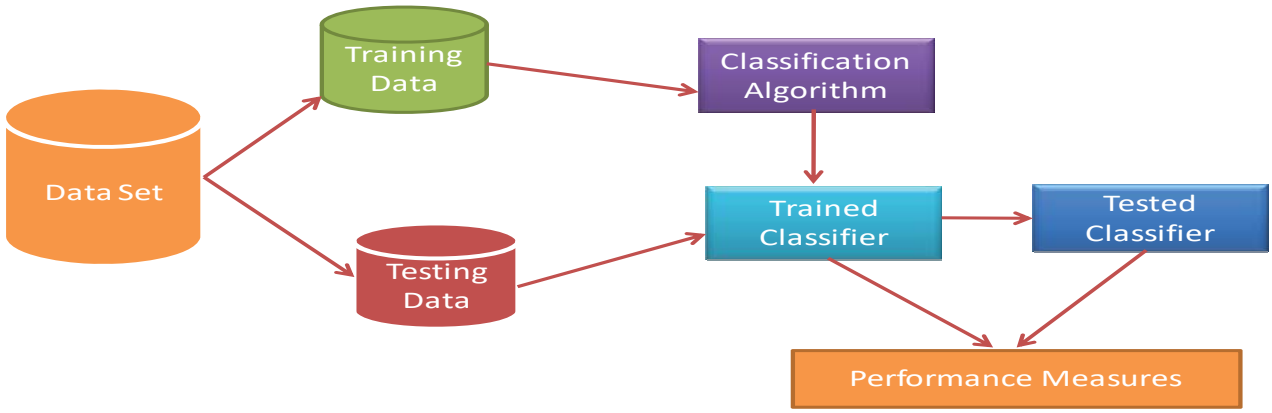

Figure 2: General Architecture of Classification

This paper focus on the classification of data that is analyzed by different authors. There are various classification techniques that can be used for various types of classification of data. Some of algorithms are define below:

- *Decision Trees-*

Decision tree [3] is a classification scheme and is used to be find description of several predefined classes and classify a data item into one of them. Tree shaped structures to predict the future trends and each branch represents set of rules for the classification of data .The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves. In a decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function

of the input attributes values.

- Iterative Dichotomiser 3 (*ID3)*-

The ID3 stands for Iterative Dichotomiser 3 algorithm beings with the original set as the root node. ID3 is the precursor to the C4.5 algorithm. It is typically used in machine learning. ID3 [3] is based on the Concept Learning System (CLS) algorithm. It is very simple decision tree algorithm.

- C4.5-

C4.5 [3] is an extension of ID3 that accounts unavailable values, continuous attribute value ranges ,pruning of decision trees and rule derivation. The decision trees generated by C4.5 can be used for classification .It is a statistical classifier. It uses gain ratio as splitting criteria.

- *Classification and Regression Trees (CART)*-

CART stands for Classification and Regression Trees. It is [3] one of the popular of building trees in the machine learning community. CART builds a binary decision tree by splitting the records at each node. CART can[4] consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution.

- *Chi-square–Automatic–Interaction–Detection (CHAID)*-

CHAID (Pujari, A. K., 2011) [4] is decision tree algorithm proposed by Hartigan in 1980. CHAID is a derivative of AID(Automatic Interaction Detection),proposed by Hartigan in 1975.CHAID attempts to stop growing the tree before overfitting occurs, where as the above algorithms generate a fully grown tree and then carry out pruning as post processing step. In that sense, CHAID avoids the pruning phase. In the standard manner, the decision tree is constructed by partition the data set into two or more data subsets, based on the values of one of the non-class attributes. After the data set is partitioned according to the chosen attributes, each subset is considered for further partitioning using the same algorithm. Each subset is partitioned without regard to any other subset. The process is repeated for each subset until some stopping criteria is met. In CHAID, the number of subsets in a partition can range from two up to the number of distinct values of the splitting attribute. In this regard, CHAID differ from CART, which always forms binary splits.

- *Quick Unbiased Efficient Statistical Tree (QUEST)*-

QUEST (*Yu-Shan, S. et al., 2011) [12]* is a binary-split decision tree algorithm for classification and data mining developed by Wei-Yin, L. (University of Wisconsin-Madison) and Yu-Shan, S. (National Chung Cheng University, Taiwan). The objective of QUEST is similar to that of the CART(TM) algorithm, Classification and Regression Trees, by Breiman, Friedman, Olshen and Stone (1984). The major differences are:
QUEST uses an unbiased variable selection technique by default.
QUEST uses imputation instead of surrogate splits to deal with missing values.

- QUEST can easily handle categorical predictor variables with many categories.

If there are no missing values in the data, QUEST can optionally use the CART Algorithm to produce a tree with univariate splits.

E. Bayesian Classification -

Bayesian classifiers [2] are statistical classifiers. They can predict class membership probabilities. Bayesian classification is based on Bayes theorem .A simple Bayesion classifier known as a native Bayesian classifier to be comparable in performance with decision tree and neutral network classifier. Bayesian classifiers have also exhibited high accuracy and speed when applied large databases. Bayesian belief network can also be used for classification.

F. Support Vector Machine (SVM) -

Support Vector Machine are learning model with associated learning algorithms that analyze data used for classification

and regression analysis. It creates a discrete hyper plane in the descriptor space of the training data and compounds are classified based on the side of hyper plane.Support Vector Machine (SVM)[11] is based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes comprises the input, making the SVM a non probabilistic binary linear classifier.

G. Artificial Neural Network (ANN)-

An artificial neural network (ANN)[3] are analytic techniques modeled after the processes of learning in the cognitive functions of the brain and capable of predicting new observation from other observation . It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning. Figure 3 shows that general architecture of artificial neural network (ANN) consists input layer, hidden layer and output layer.

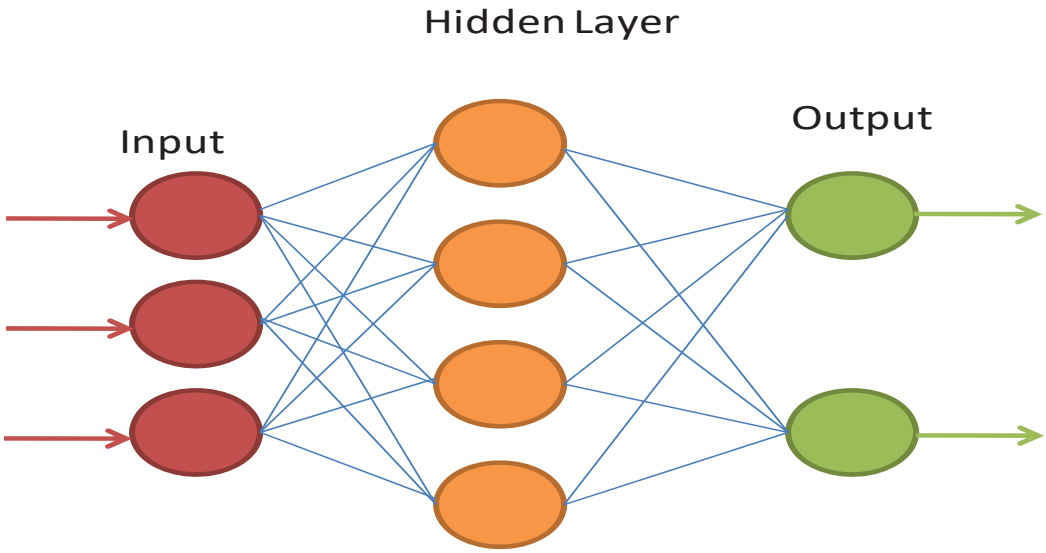

Figure 3: General Architecture of ANN

V. CONCLUSION

Teaching assistant evaluation is one of the important criteria in field of education sector. In academic sector, to improve the education quality, teaching activities are major and important factors. This research work focus on the applications of data mining techniques that is applied on teaching assistant evaluation. We have mainly described about various classification techniques in the context of teaching assistant evaluation.

REFERENCES

[1] J. Yang, H. Jiang and H. Zhang, "Teaching Assistant Evaluation Based on Support Vector Machines with Parameters Optimization", Information Technology Journal, Vol.No. 10, pp. 2140-2146 , 2011.

[2] J. Han and M. Kamber,"Data Mining: Concept and techniques" Simon Fraser University,2005.

[3] B. B. Agarwal and S. Prakash, "Data Mining and Data Warehousing" I.F.T.M. University Moradabad (U.P.) UDM-9324-28,2009.

[4] A. K. Pujari

[5] A. Gupta, "Classification Of Complex UCI Datasets Using Machine Learning And Evolutionary Algorithms" , International journal of Scientific and technology research,Vol.No.4 , pp. 94 , 2015.

[6] E.G. Dragomir, "Teaching Performance Evaluation Using Supervised Machine Learning Techniques",The 5th International Conference on Virtual Learning (ICVL) ,pp. 390-394, 2010.

[7] A. Alias Balamurugan Subramanian, Spramala, B. Rajalakshmi, R. Rajaram "Improving Decision Tree Performance by Exception Handling", International Journal of Automation and Computing , Vol. 7(3), pp. 372-380 ,2010 .

[8] S. Agarwal, G. N. Pandey, and M. D. Tiwari , "Data Mining in Education: Data Classification and Decision Tree Approach", Vol. No.2 , pp. 140-144. , 2012.

[9] P. Kaur , M. Singh and G. Singh Josan , "Classification and Prediction Based Data Mining Algorithms to predict slow learners in Education Sector",3rd International Conference On Recent Trends in computing , pp. , 500-508, 2015 .

[10] T. Z. Mohammad, A. M.Mahmoud, El-Sayed M. El-Horbart, Mohamed I.Roushdy and Abdel-Badeeh M. Salem, "An Intelligent Educational Data Mining Classification Model for Teaching English for Slow Learner Students" , IPASJ International Journal of Computer Science ,Vol. No. 2 ,pp. .6-15.

[11] L. Jena, and N. K. Kamila , "Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease", International Journal of Emerging Research in Management &Technology ,Vol.No.-4,pp. 110-118 ,2015.

[12] *Yu-Shan S. and Wei-Yin L. (2011), QUEST Classification Tree (Version 1.9.2),* http://www.stat.wisc.edu/~loh/quest.html . *(Browsing date: 25th April 2012).*

[13] *S. Mardikyan, B. Badur,"* Analyzing Teaching Performance of InstructorsUsing Data Mining Techniques", Informatics in Education, · 2011 *Vilnius University, pp. 245-257,2011.*