Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparing the Applications of Various Algorithms of Classification Technique of Data Mining in an Indian University to Uncover Hidden Patterns

Sohil Pandya
MCA Department,
Sardar Vallabhbhai Patel Institute of Technology (SVIT),
Vasad – 388306
Gujarat, India

Dr. Paresh V. Virparia
G H Patel PG Dept of Computer Science & Technology,
Sardar Patel University,
Vallabh Vidyanagar – 388120,
Gujarat, India

Abstract—Education Institutes or Universities can sustain and become centre of excellence if they are able to do effective and efficient data analysis. The application of various algorithms of Classification Technique of Data mining in an Indian University is an attempt to uncover hidden patterns along with measuring a fitment of the algorithms in the domain. The experimental results are enough motivating to researchers for selection of the algorithm for uncovering hidden patterns by various parameters.

Keywords— Classification, Decision Tree, Feed Forward Neural Networks, Naïve Bayes, Support Vector Machine

I. INTRODUCTION

Educational Institutes or Universities are the backbones for the countries. They are ultimately play vital role for the growth of the countries. In this era of globalization, they can sustain if they are having competitive advantage [8]. Efficient and effective data analysis of the data injected is one of the ways to achieve competitive advantage. These injected data of students enrolments, results, etc are very much useful for developing models for students' selection of discipline, performance, required infrastructure, syllabus updating like strategic decisions [1, 2, 12, 13]. Data Mining is way to uncover hidden patterns from the large databases for better data analysis.

There are several techniques of data mining like clustering, association rule mining, classification, outlier analysis, etc. for uncovering hidden patterns from this data [5, 6, 7]. There are various algorithms of above techniques are developed by various researchers. In this paper researchers tried to examine and investigate various methods of classifications like Decision Trees (DT), Naïve Bayes (NB), k-Nearest Neighbour (kNN), Feed Forward Neural Networks (FFNN) and Support Vector Machine (SVM) to identify the best fit methods among them for the University domain. All the above mentioned algorithms were implemented using WEKA, an Open Source Software which consists of a collection of machine learning algorithms for data mining tasks.

In the next section, brief introduction to all above classifiers are mentioned followed by experimental results and discussions. The result of the study will be useful for the research to find out fitness of the techniques and researchers who wish to determine the technique for application of data mining in uncovering hidden patterns.

II. CLASSIFIERS

In this section, the general an introduction to each of the selected classification algorithms are given as follows:

A. Decision Trees

Decision tree induction is the learning of decision trees from class-labelled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an out come of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node [16].

B. k-Nearest Neighbour

In pattern recognition, the k-nearest neighbour algorithm (k-NN) is a non-parametric method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbour [14].

C. Naïve Bayes

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model" [14].

D. Feed Forward Neural Network

A feed forward neural network is a biologically inspired classification algorithm. It consists of a (possibly large) number of simple neuron-like processing units, organized in layers. Every unit in a layer is connected with all the units in the previous layer. These connections are not all equal, each connection may have a different strength or weight. The weights on these connections encode the knowledge of a network. Often the units in a neural network are also called nodes. Data enters at the inputs and passes through the network, layer by layer, until it arrives at the outputs. During normal operation, that is when it acts as a classifier, there is no feedback between layers. This is why they are called feed forward neural networks [15].

E. Support Vector Machines

Support Vector Machines (SVM) are supervised learning models with associated learning algorithm that analyses data and recognize patterns, used for classification and regression analysis [14]. It takes a set of input data and predicts, for each given input which of two possible classes forms output, making it a non-probabilistic binary linear classifier.

III. METHODOLOGY

The obtained data is pre-processed with according to the need of the system. In particular region field is derived field from the district. So the district field was cleaned using Context Free Data Cleaning [4, 9, 10]. The numerical values of Percentage are converted into discrete values of Grade. At the end the entire dataset is converted into .CSV format. It is then supplied to WEKA.

The following classification algorithms were implemented in WEKA:

- Decision trees were obtained in WEKA J48 algorithm (weka.classifiers.trees.j48) (Java implementation of C4.5 algorithm) with Support 0.5 is used.
- k-Nearest Neighbour using weka.classifiers.lazy.lbk
- Feed Forward Neural Network using weka.classifiers.functions.MultilayerPerception
- Support Vector Machine using weka.classifiers.functions.SMO
- Naïve Bayes using weka.classifiers.bayes.NaiveBayes

All the results of above algorithms are summarized and explained in next section.

IV. EXPERIMENTAL RESULTS & DISCUSSIONS

The experimental results, as depicted in Table I and subsequent charts, were quite motivating and leading to the fulfilment of objective.

TABLE I
COMPARISON OF VARIOUS CLASSIFIERS

Parameters / Classifiers	% of Correctly Classified Instances	Kappa Statistics	Avg. ROC Area	Mean Absolute Error
DT	63.09	0.24	0.69	0.17
kNN	63.32	0.26	0.77	0.17
FFNN	63.27	0.26	0.77	0.16
SVM	62.78	0.25	0.68	0.24
NB	62.04	0.22	0.76	0.17

The experimental results can be interpreted as follows:

(i) FFNN and kNN are classifying the dataset more accurately compare to others.
(ii) Kappa statistics is used to asses the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of data collected and their validity [4, 17]. Out of the all the above algorithm the FFNN and kNN are more significant.
(iii) ROC (Receiver Operating Characteristics) Area is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the faction of false positive out of the negatives. The average ROC of FFNN and kNN are found similar, fair and highest compare to others. [14]
(iv) Also the error rate in FFNN is lowest compare to others.

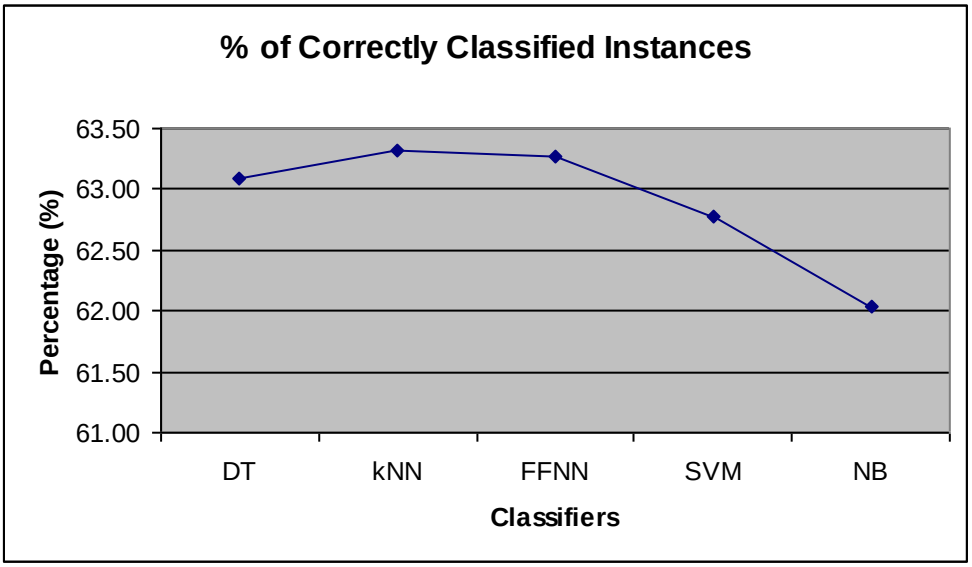

Fig. 1 Chart for % of Correctly Classified Instances

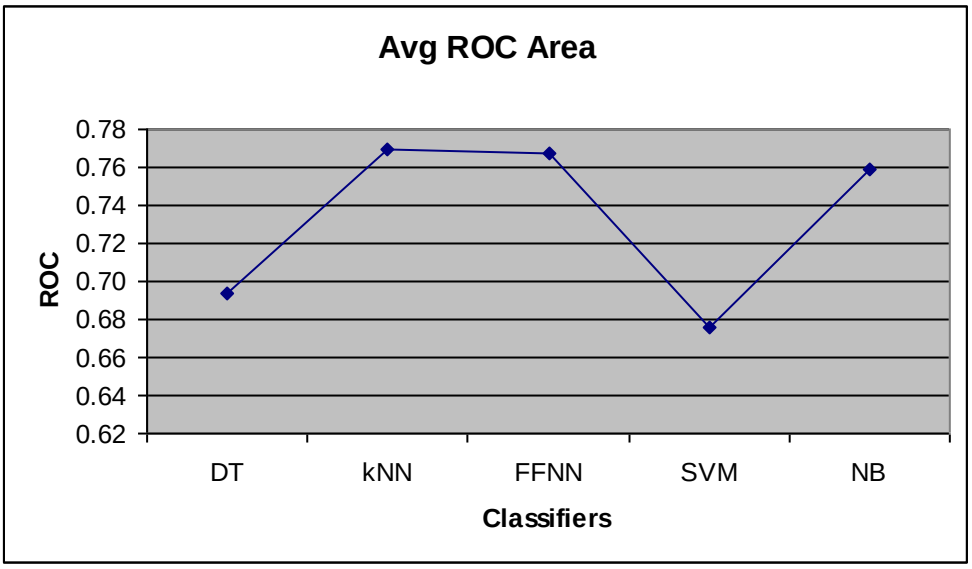

Fig. 2 Chart for Average ROC Area

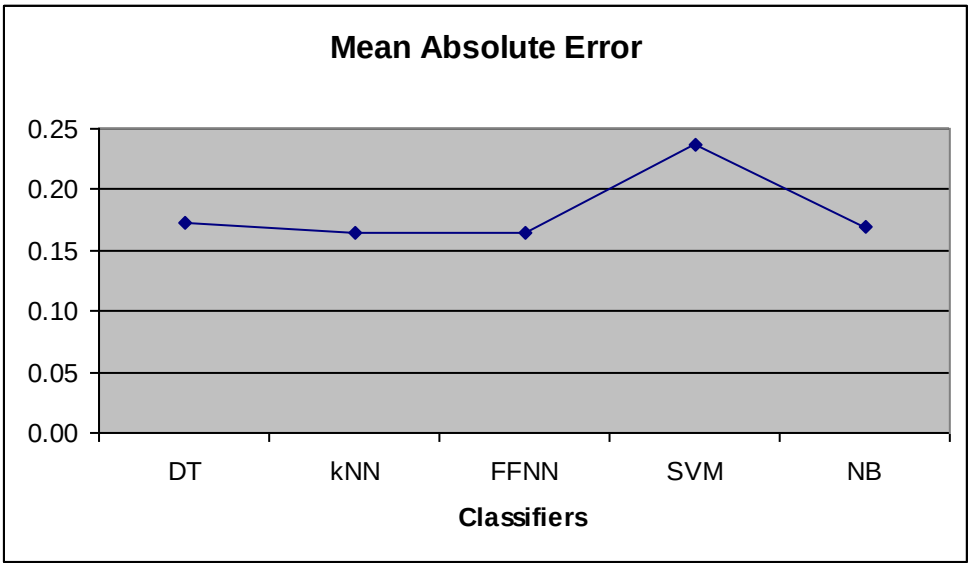

Fig. 3 Chart for Mean Absolute Error

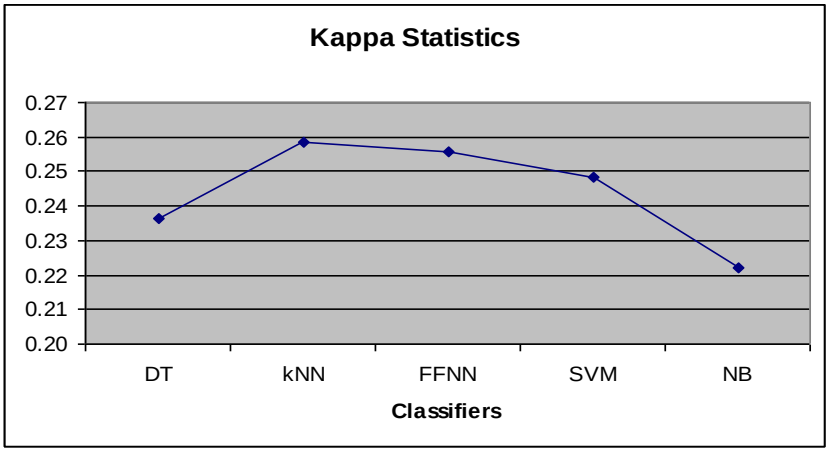

Fig. 4 Chart for Kappa Statistics

V. CONCLUSION

With a view of above results and discussion, we met our objective to investigate and evaluate five selected classification algorithms using WEKA. The fit algorithm with various measurable parameters is found Feed Forward Neural Networks. Support Vector Machines (SVM) Algorithm is found as a least fit algorithm. Hence, the Feed Forward Neural Network algorithm has potential to notably pick up the conventional classification methods for use in University or in general in Educational Data Mining. The experiments done above can be replicated as and when needed with different inputs to uncover other patterns. The experiments done above with available and limited data set. The results may differ if the different data set is used or another strategy is applied, which may be tested. There may be other determinants which may not be present in the dataset or may be overlooked by the authors while experimenting. Based upon the specific dataset, with above framework and methodology we can have more specified results, which can be used for various strategic decisions.

REFERENCES

[1] Ashutosh Nandeshwar, Subodh Chaudhary, "Enrollment Prediction Models using Data Mining", retrieved on April 10, 2012 from http://nandeshwar.info/wpcontent/uploads/2008/11/DMWVU_Project.pdf.

[2] B. K. Baradwaj, S. Pal, "Mining Educational Data to Analyze Students' Performance", in International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[3] Hui Xiong, Gaurav Pandey, Michael Steinbach, Vipin Kumar "Enhancing Data Analysis with Noise Removal" in IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 3, pp. 304-319, March 2006

[4] Mohd Fauzi bin Othman, Thomas Moh Shan Yau, "Comparison of Classification Techniques using WEKA for Breast Cancer" in Biomed 06, IFMBE Proceedings, pp. 520-523, 2007

[5] Oladipupo O O, Oyelade O J, "Knowledge Discovery from Students' Result Repository: Association Rule Mining Approach", in International Journal of Computer Science & Security, Vol 4, No 2, pp. 199-207, November 2009

[6] R R Kabra, R S Bichkar; "Performance Prediction of Engineering Students using Decision Trees" in International Journal of Computer Applications, Vol 36 No 11, pp. 8-12, December 2011.

[7] S. Anupam Kumar, Vijayalaxmi M N; "Efficiency of Decision Trees in Predicting Student's Academic Performance", in International Journal of Computer Science & Information Technology (CS&IT), Vol 1, No 2, pp. 335-343, 2011.

[8] S Numprasertchai, Y Poovaravan, "Enhancing University Competitiveness through ICT Based Knowledge Management Systems", in Proc. of IEEE Int. Conf. on Management of Innovation & Technology, Volume - 1, pp. 417-421, June 2006.

[9] Sohil D. Pandya, Dr. Paresh V. Virparia, "Studying in impact of Past Performance in Academics using Data Mining Technique", in Inter-National Journal of Information and Computing Technology, Vol 2 No 1, January 2012

[10] Sohil D. Pandya, Dr. Paresh V. Virparia, "Context Free Data Cleaning and its Application in Mechanism for Suggestive Data Cleaning" published in Inter-National Journal of Information Science, Vol 1 No 1, February 2012

[11] V. P. Bresfelean, "Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment", in Proc. of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, June 25-28, 2007.

[12] Zlatko J Kovacic, "Early Prediction of Students Success: Mining Students Enrollment Data", in Proceedings of Information Science & IT (InSITE), 2010.

[13] Zlatko J. Kovacic; "Predictive working tool for early identification of 'at risk' students", published in Creative Commons 3.0 New Zealand Attribution Noncommercial Share Alike License.

[14] http://en.wikipedia.org/

[15] http://www.fon.hum.uva.nl/

[16] Data Mining Concepts and Techniques, Jiawei Han and Micheline Kamber, 2nd Edition, Elsevier

Kappa at http://www.dmi.columbia.edu/homepages/chuangj/kappa

