DOI 10.29042/2017-1984-1987

Helix Vol. 7(5): 1984-1987

Suspicious Defaulter Forecasting Using Machine Learning Classifiers

[*1]Ashmeet Singh, [2]Pranav Sharma, [3] Prof. L. Shalini

[1, 2, 3] VIT University, Vellore, India

Email: [1] ashmeet.singh2013@vit.ac.in,[2] pranav.sharma2013@vit.ac.in, [3]lshalini@vit.ac.in

Received: 4th May 2017, Accepted: 15th June 2017, Published: 1st September 2017

Abstract

Every year individuals join certain University College filled with diverse set of people which deems to be a stepping stone towards the road one wants to go on. The aim of any University/ College is to create an environment for individuals to study and succeed in their life. Achieving this involves various activities. One of the most important involved in this process is to identify and guide the defaulters to the path of success. Identification of suspicious defaulters among thousands of individuals is a very big task itself. Such a problem holds the key to the success or downfall of the University/ College. Survey was carried out to collect large amount of data for analysis. On manual analysing the given data 23 factors were found to be contributing towards the result. Then the data was further analysed through five classification algorithms to check the result through various aspects. Among the various algorithms, the most accurate algorithm, Decision Tree, was selected for implementation purpose. The algorithms provided the result in form of ranking as ok, doubtful and very doubtful. Every student must be categorized into one of these ranks and certain set of actions associated to them could be carried by the management to maintain discipline in campus.

Decision Tree algorithm is further coded in python. The python coding act as a backend for the front end of the created web application. Frontend web application provides with an individual suspicion and list of suspicious defaulter which is found by analysis and validation of individual data. The security is maintained through a login page. Such solution ease out the process of suspicion for the College/ University and makes the process faster. It also helps them to guide a niche crowd and not as a whole. This helps the University/ College both qualitatively and quantitatively.

Keywords—Decision Tree, RapidMiner, Naive Bayes, Random Forests, K-NN, Neural Networks.

Introduction

For ensuring a good reputation of the university, a quick identification of defaulters is required. Currently the identification of the defaulters is operated in a non-digital or semi-digital manner. With the rise of machine learning techniques, a technical solution has to be proposed to make the manual process faster, niche, effective, productive and automated.

The proposed system is a web application that views database of defaulters. This list is generated by a prediction algorithm that acts best among all algorithms for the desired dataset. The parameters, such as accuracy, kappa and F-measure, ensures an unbiased and a clear judgement over the classifiers. The dataset was created using the survey of students and close inspection of experts. The data represents the dependencies of various aspects of student details for being involved in forbidden activities. The main focus of discussion will be the selection of best classifier with appropriate tests and analysis conducted on rapidminer tool. The best classifier will be further be used in implementation of the web application. The developed application can play a vital control in easing down the management process. Students will be compelled to maintain an appropriate rank along with the academic grades. Section II gives the overview of previous related works. The detailed explanation of Dataset and its attributes is given in Section III. Short discussion and the approaches of all the classifiers are given in Section IV. In next section, experiments and analysis are presented and then followed by results and conclusion of the discussed approach.

Literature Survey

The main aspects for the project are based on the survey work done by two scholars of University of Minho (Paulo Cortez and Alice Silva) presenting the views on predicting secondary school student performance. They focused on improving quality of education and boosting management procedures of school. The data collected by survey was analyzed with two classifiers (Decision Tree and Random Forest) and a regression technique. The Grading scheme was adopted and was awarded according to the results and performances of the students. There main aim was to focus on a new grading system which will reflect the social values such as discipline, punctuality, avocation tasks (such as alcohol consumption, reading books etc.) etc.[5]

Moreover, when it comes to classification and prediction algorithms, the testing must go on with other contending classifiers such as Naive Bayes and Neural Networks. Therefore, with the study of different algorithms and their respective specialties were studied under different research papers and five among them were shortlisted apparently as the best algorithms for our datasets. [12]

Given in the above survey, the results were

limited to three prediction algorithms and attribute selection were not as appropriate. As 'Suspicious Defaulter Forecasting' project is based for a university scenario (especially universities having detailed database of students keeping in mind availability and feasibility of student data as one in VIT University), some of the attributes must be ignored from the above survey (example age of college student is greater than that of a school student). To achieve the implementation of the system, some attributes have been added which are available by university database and could play a crucial role in judgement (for example the university has a record of fines paid by student).

In addition, the implementation of system and its description are missing in project done by Paulo and Alice which are overcomed by our project [5].

Materials and Methods

Dataset

Dataset used consists of 21 feature attributes, a target attribute and an identification attribute. The weightage of each feature attribute differs and could be given by gini index. Alcohol attribute has the highest gini value followed by others with 'FSize' having the least one. The dataset consists of over thousand instances. Abbreviation of the attributes are explained in the following table.

Table 1: Dataset related variables

Gender	**student's gender (binary: male or female)**
Age	student's age (numeric: from 17 to 28)
Address	student's hometown category (binary: urban or rural)
FSize	student's family size (<=3 or >3)
ParStat	student's parents coexistence status (binary: together or apart)
MotherEdu	mother's literacy (numeric: from 0 to 4)
FatherEdu	father's literacy (numeric: from 0 to 4)
MotherJob	student's mother profession
FatherJob	student's father profession
AdmSeat	student's admission seat type (binary: merit or management)
Guardian	student's guardian (binary: mother or father)
StudentType	student's type (numeric: hosteller, day scholar or day scholar with guardian)
Fineamount	Number of times fine paid (numeric: 0 to 5)
Arrear count	Number of Arrear (numeric: 0 to 5)
Activities	Student involved in co - curricular (binary: yes or no)
Counselling	Student taking counselling (binary: yes or no)
Home Leave	Student's no of leave to home (numeric: 0-10)
Alcohol	whether student drink alcohol or not (binary: yes or no)
Health	student's health factor (numeric: 1-6)
Absences	students attendance 75% above or not (binary: yes or no)
cgpa	students cgpa (numeric : A(90-100), B(80-90), C(70-80), D(60-70), E(50-60))
Suspicious	student's suspicious factor (numeric: very doubtful, doubtful or ok)
ID	student's ID

Classifiers

A. Naive Bayes

Naive Bayes is a probabilistic classification which model the probability of class memberships given the input data and make a probabilistic model of data of each class. Here all variables of each class are independent of each other. Probability of each class is estimated and the class with maximum probability is choosen.It is best applied in filtering spam messages [7].

B. Decision Tree

Decision Tree looks at attribute of every class and tries to give out the best inference. Here the algorithm will look for the attributes in the data and use those attributes to split the data into subsets. If the subset is pure, it stops or else it keeps on splitting. When a new dataset is encountered, a check will be made to see which subset example falls into and use the dominant class in that subset [8].

C. Random Forests

Random Forest works as a large collection of decorrelated decision trees. It creates a lot of decision trees and use them for classification. Here a

lot of subset is created based with random values i.e decision tree. Thus obtained are different variation of main classification used to create ranking of different classifier [9].

D. Neural Networks (NN)

Neural network is like any other network consisting interconnected web of nodes called neurons and the edges that join them together. They are used for classification of tasks where an object can fall into one of at least two different categories. It is structured and comes in input and output layer and all between these layers are termed as hidden layers. It receives a set of inputs, performs progressively complex calculations and then use the output to solve a problem [10].

E. K- Nearest Neighbours (K-NN).

K-Nearest Neighbours identifies the nearest neighbors of the element that defines the class. Given N training vectors, K-NN identifies the k nearest neighbors of 'c', regardless of labels [11].

Analysis of the 5 different algorithms of classification known as 'Classifiers' is carried out and judged on 5 parameters named as Accuracy, Class Precision, Class Recall, Kappa Measure and F Value. The five parameters used for judgement are as follows [8]:

- Accuracy = ((True Positive + True Negative) / (P + N))*100
- Precision = (True Positive / (True Positive + False Positive))*100
- Recall = (True Positive / (True Positive + False Negative))*100
- F-Measure = (2*Precision*Recall / (Precision+ Recall))*100.
- Kappa = (observed accuracy - expected accuracy)/ (1 - expected accuracy)

Experiment and Discussions

All five classifiers are made to run on Rapidminer tool for finding out who does the best job among them. Following results and observations were noted for the respective classifiers

A. Naive Bayes

Simple probability based classifier is good enough to act upon small datasets and gives accuracy of prediction near to 89%. It is very fast and does classification in no time.

B. Decision Tree

The algorithm is based on the calculating the most desirable attribute on which the label attribute is dependent. It is done by calculating Gini Index and Gain Ratio. It acts best for medium sized datasets and gives the highest accuracy of 90% along with F-measure as 0.811.

C. Random Forest

Since Random Forest is suitable on very large datasets therefore randomness created while generating the tree doesn't helps the desired dataset of 1045 instances.

D. NN

Neural Network being the fuzzy classifier gives the optimum results similar to Naive Bayes but the amount of time taken for the execution takes away all its efforts. The time taken was more than 3 minutes which more than 200 hundred times slower when compared to other classifiers.

E. K-NN

K-NN being the clustering classifier doesn't meet our requirements therefore due to lack of kappa measure its uncertainty pulls it down for further proceedings.

	Accuracy	**Precision**	**Recall**	**F-measure**	**Kappa**
Naive Bayes	89	84.5	73.6	78.7	69
Decision Tree	**90***	**89***	**74.52***	**81.1***	**76***
Random Forest	79	82.4	41.5	55.2	13
Neural Network	87	81.6	78.5	80	68
K-NN	78	66.1	63.8	64.9	45

Table 2: Comparison results of different classifiers.
* Maximum Value

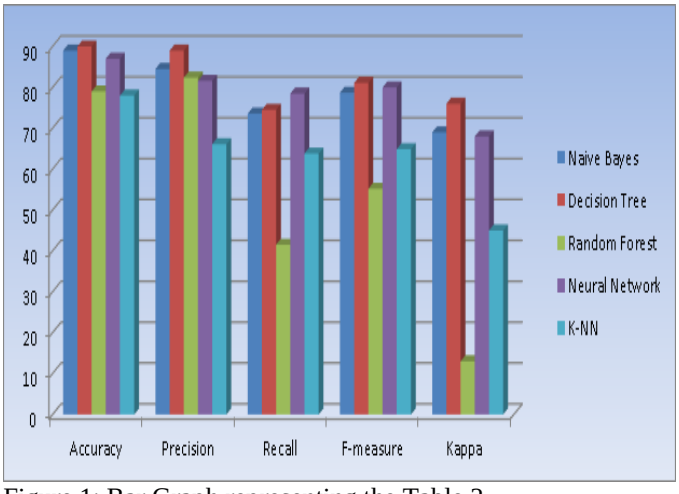

Figure 1: Bar Graph representing the Table 2.

Results and Discussion

The survey, carried out in Rapid Miner tool over the dataset acquired by surveys and close inspections by experts, gives highest accuracy over decision tree algorithm. Even though Naive Bayes is highly

competitive but with bit low kappa value it results on second place. As the accuracies of both Random Forest and K-NN are comparable but with high variations of random forest over small dataset makes it least appropriate for use.

Conclusion, Limitations and Future Scope

The system supports the adaption of new system which makes the management process easier. A new competitive approach could be build which will enhance and motivate students to be morally and socially respectable.

Feature selection could further be improved and different strategies could be involved for more influencing attribute generation (e.g. Medical checkup every month for better accuracy of prediction).

The system gives the rudimentary steps towards the future grading system for keeping check over the students. In future, the process could be made as the grade awarding process by applying grade values (A+, A, B+, B, D) by further distribution of values (Okay, Doubtful and Very doubtful) of label attribute (Defaulter list).

Acknowledgment

Would like to thank our guide Prof. L. Shalini (Assistant Professor – Senior) who provided insight and expertise that greatly assisted the research. Thanks to R. Senthil Kumar (Head, B.Tech (CSE), SCOPE, and VIT University.) for assistance with research and for comments that greatly improved the manuscript. Further would like to show our gratitude to Prof. Arunkumar T (Dean, SCOPE, and VIT University) for sharing his pearls of wisdom with us during the course of this research, and thank our “anonymous” reviewers for their insights and for the comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons. This research was supported/partially supported by VIT University, Vellore.

References

[1] N. Friedman, D. Geiger, and Goldszmidt M. Bayesian network classifiers. Machine Learning, 29:131–163, 1997

[2] R. Diaz-Uriarte and S.A. de Andres. Gene selection and classification of microarray data using ´ random forest. BMC Bioinformatics, 7:1471–2105, 2006.

[3] N. Murata, S. Yoshizawa, and S. Amari, ―Learning curves, model selection and complexity of neural networks,‖ in Advances in Neural Information Processing Systems 5, S. Jose Hanson, J. D. Cowan, and C. Lee Giles, ed. San Mateo, CA: Morgan Kaufmann, 1993, pp. 607-614

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, pages 1106–1114, 2012.

[5] Cortez, Paulo, and Alice Silva. "Using Data Mining to Predict Secondary School Student Performance". Web. 9 Oct. 2004.

[6] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood," Random Forests and Decision Trees", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[7] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam," Comparative Study of Classification Algorithms used in Sentiment Analysis", International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014

[8] Shahrukh Teli, Prashasti Kanikar " A Survey on Decision Tree Based Approaches in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015

[9] S.L. Ting, W.H. IP, Albert H.C. Tsang "Is Naïve Bayes a Good Classifier for Document Classification?", International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, 2011.

[10] Puyalnithi, Thendral, Madhu Viswanatham V, and Ashmeet Singh. "Comparison of Performance Of Various Data Classification Algorithms With Ensemble Methods Using RAPIDMINER". International Journal of Advanced Research in Computer Science and Software Engineering 6.5 (2016): n. pag. Print.

[11] Ashmeet Singh and Ananda Kumar S. "Application of Bagging and Boosting for all the Classification Algorithms". International Journal of Pharmacy & Technology 8.3 (2016): n. pag. Print.

[12] Ashmeet Singh and R Sathyaraj. "A Comparison between Classification Algorithms On Different Datasets Methodologies Using Rapidminer". International Journal of Advanced Research in Computer and Communication Engineering 5.5 (2016): 4. Print.

  ISSN 2319 – 5592 (Online)