A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining

Keshav Singh Rawat and I. V. Malhan

Abstract Machine learning algorithm can be applied in education data mining (EDM) to extract knowledge. Educational data mining is an important practice of automatic extraction and segmentation of useful information from the education data sources. This paper is focused on comparison and study of hybrid model of classification and machine learning algorithms based on decision tree, clustering, artificial neural network, Naïve Bayes, etc. This paper introduces concepts of popular algorithm for new researchers of this area. The paper discusses hybrid classification model using machine learning algorithms using voting that can be used to analyze the performance of students. We have used open source data mining tool Weka for a practical experiment on data set of students that serve the purpose of prediction, classification, visualization, etc. The findings of this paper reveal that hybrid method of classification are more efficient for prediction of student-related data.

Keywords Machine learning · Educational data mining · Decision tree · Hybrid classification · NB · ANN

1 Introduction

Data mining is a process of automatics extraction of useful information from the large data sets repositories, etc. Data mining helps to find relationships and discover patterns by using machine learning algorithm, statistics, and visualization. Data mining application in the field of education facilitates better discovery of the learning resources, and improves efficiency of learning process. The main objective of educational data mining is to improve the learning quality and understanding that helps us

K. S. Rawat (✉) · I. V. Malhan
Department of Computer Science & Informatics, Central University of Himachal Pradesh, Dharamshala, Himachal Pradesh, India
e-mail: keshav79699@gmail.com

I. V. Malhan
e-mail: imalhan_47@rediffmail.com

C. R. Krishna et al. (eds.), *Proceedings of 2nd International Conference on Communication, Computing and Networking*, Lecture Notes in Networks and Systems 46, https://doi.org/10.1007/978-981-13-1217-5_67

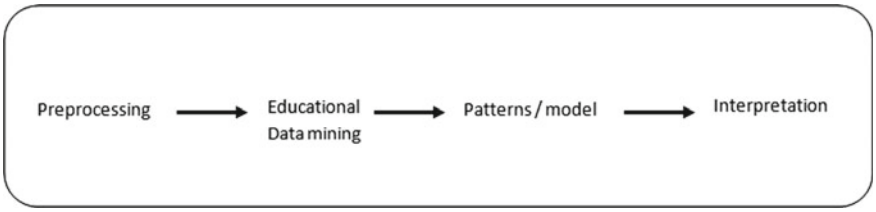

Fig. 1 Process of educational data mining (EDM)

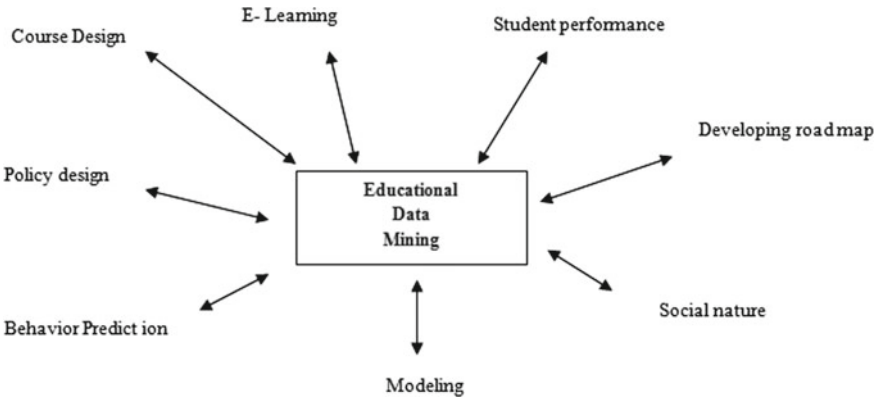

Fig. 2 Process of educational data mining (EDM)

to improve learning as well as strategies designing for administrative planning [1]. In this paper, we discuss machine learning algorithms applied on the student data set. These machine learning algorithms are capable to extract useful information from education-related data sets.

The process of educational data mining (EDM) is described in Fig. 1 [2, 3]. EDM contains mainly four stages. In the first phase, we discussed relationship between data, the second phase contains validation process, the third phase predicts result and the final phase is used for decision-making process. Based on the output of the third phase, the EDM is applied in some areas [4] described in Fig. 2.

The organization structure of the paper is as follows. Section 2 contains some of the interrelated work in the literature review. Section 3 discusses the machine learning algorithm used in this study. Section 4 describes the proposed work. Section 5 shows experiment and results of this paper. Section 6 concludes the work and final section contain references.

2 Literature Review

Many papers have been published in the area of educational data mining. Data mining in education field is very useful to extract useful information of educational outcomes that helps to administrators to make policies regarding learning and performance of students. The author of paper [5] discussed and reviewed various applications, tools, and concepts of EDM. The applications and tasks of educational data mining are discussed in detail in [6] and reviewed educational data mining tasks and decision support system to improve the performance of student. The research in the field of educational data mining (EDM) and educational learning analysis improved the overall education system and decision support system [7]. The hybrid classification approach [8] was described method of ensemble various machine learning algorithms to achieve better performance in term of precision, recall, and F measure on various data sets.

3 Machine Learning Algorithms

Machine learning is a method of producing algorithm to predict useful information from source data set. Machine learning algorithms can be applied in supervised or unsupervised way. The process of machine learning contains three basic steps—Initialization, learning and testing that provide efficient and accurate output data. This field can be applied on many application areas—face identification, weather prediction, speech reorganization, classification, fraud detection, spam filtering, and so on. In this paper, we discuss some important machine learning algorithms of data mining and implementation of these algorithms to estimate the result of students. The four machine learning technique—decision tree, Naïve Bayes, multilayer perception, and K-nearest neighbor algorithms have been used on student data set in this paper.

3.1 Decision Tree (J48)

Quinlan [9] developed decision tree algorithms ID3 and C4.5. In Weka, C4.5 is known as J48 algorithm. This tree method based on divide and conquer approach to generate decision tree by recursive calls. C4.5 algorithm is an advanced algorithm as compared with ID3 and overcomes all difficulties and limitations of ID3. This algorithm is used gain ration as feature selection of an attribute to create decision tree. The attribute with highest gain ratio holds the position of parent node and next values hold the position of children and so on till the completion of decision tree. There are two ways to create tree, pruned, or unpruned. The pruning method helps to improve the efficiency of decision tree by removing unwanted branches. Figure 3 shows the decision tree on student data set using J48 algorithm in Weka.

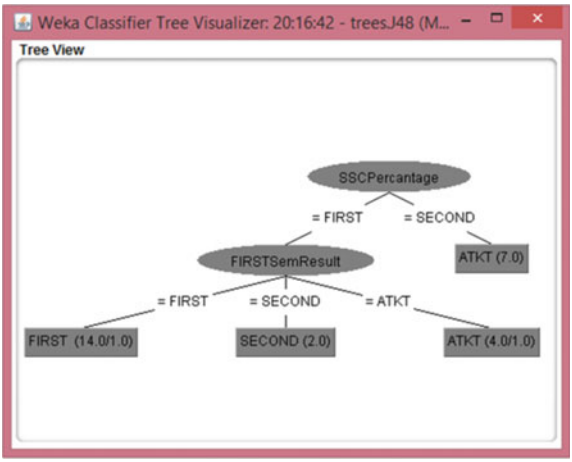

Fig. 3 Decision tree structure

3.2 K-nearest Neighbor Algorithm (IBK)

The K-nearest neighbor algorithm for classification in Weka is known as IBK algorithm comes under lazy learning algorithms. IBK algorithm uses Euclidean distance to measure K-nearest neighbor for classification data set.

3.3 Naïve Bayes (NB)

It is the popular classification technique based on conditional probabilities and Bayes theorem [10]. Naïve Bayes classification method is very useful for text classification. It takes less training data for classification of data.

3.4 Multilayer Perception (ANN)

In Weka, artificial neural network concept is offered by multilayer perception (MLP). This method contains an input layer, an output layer, and some hidden layers and used backpropagation approach for classification.

4 Proposed Method

The decision tree (J48), Naïve Bayes (NB), K-nearest neighbor (IBK) and multi-perception (ANN) machine learning algorithms are used to develop hybrid classifi-

cation model using voting method. The detailed description of method is reflected in Algorithm 1. Here, all four algorithms mentioned above are grouped using voting approach to get better efficiency of prediction. A 10-fold cross-validation is used to predict the evaluation accuracy.

Algorithm1: Proposed work
I/P: Student data set for classification
O/P: Performance of classification in terms of accuracy.

Steps:

1. Input data set of students.
2. Apply preprocessing to remove unwanted information and attributes have low information gain.
3. Apply machine learning algorithms decision tree (J48), Naïve Bayes (NB), K- Nearest Neighbor (IBK), Multilayer perception (ANN) to student data set independently.
4. Development of hybrid classification model-
   (i) Find classification hypothesis of J48, IBK, ANN and NB using ensemble.
   (ii) Perform voting process by combining classifiers using average posterior probabilities rule.
5. Compare machine learning algorithm J48, NB, IBK and ANN with hybrid classification in terms of correctly classified instance, incorrectly classified instances and accuracy.
6. End.

5 Experiment and Result

This section contains implementation of machine learning algorithm on data set of students to analyze results. We used open-source Weka tools for the experiment purpose. Weka tool supports various data mining techniques and their algorithms to classify given data set. We used student data set of Department of Computer Science in csv format for classification and prediction. Data set can be easily transformed in arff format by ARFF tool provided by Weka. The student data set contains both academic and personal data. The attributes of data set are—Roll no, Address, Father's occupation, Mother's occupation, Father's Education, Gender, Caste, SSC Result, Graduation Result, First Semester Result, and Second Semester Result. The student data set contains around 27 instances and 11 attributes. The actual student data was in a continuous form that was converted into nominal form for experiment analysis. Accuracy rate is used to evaluate model of classification. The correct and incorrect classified attributes are defined by confusion matrix. In Confusion matrix, true positive and false positive rate are used to evaluate the classification model. The general confusion matrix is shown in Fig. 4, where rows indicate actual class and columns indicate prediction class. TP and TN represent the total number of true

TP: True Positive (number of positive instances correctly classified)
TN: True Negative (number of negative instances correctly classified)
FP: False Positive (number of negative instances incorrectly classified)
FN: False Negative (number of positive instances incorrectly classified)

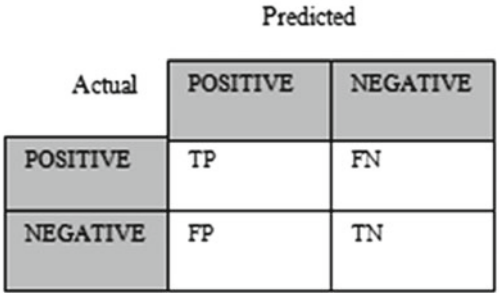

Fig. 4 Confusion matrix of binary class

positive correctly classified instance and the total number of true negative incorrectly classified instance. Accuracy of classifier is defined as the total number of correctly classified instances.

The overall performance [11] of classification can be measured by the following parameters:

Accuracy of classification is defined as the total number of instances that are correctly classified.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Table 1 show comparisons of classifiers efficiency in terms of correctly classified, incorrectly classified between proposed hybrid model and individual classifiers, i.e., J48, NB, IBK, and ANN. It is observed that the proposed hybrid method of classification achieved highest correctly classified instance over individual classifiers. Table 2 shows comparisons of accuracy or precision between proposed hybrid model and individual classifiers, i.e., J48, NB, IBK, and ANN.

It is observed that the proposed hybrid method of classification as shown in Fig. 5 is more accurate than individual classifiers, i.e., J48, NB, IBK, and ANN. This hybrid method achieved the highest accuracy of 92.59% and individual classifiers, i.e., J48, NB, IBK, ANN achieved an accuracy of 85.18, 81.48, 88.88, and 88.88% respectively.

Table 1 Comparisons of classifiers efficiency

Classifier	Execution time (s)	Efficiency (%)	Correctly classified instances	Incorrectly classified instances
Decision tree (J48)	0.41	85.18	23	4
K-nearest neighbor algorithm (IBK)	0	88.88	24	3
Naïve Bayes (NB)	0	81.48	22	5
Multilayer perception (ANN)	0.13	88.88	24	3
Proposed hybrid method	0.03	92.59	25	2

Table 2 Comparisons of classifiers accuracy

Classifier class	Decision tree (J48)	K-nearest neighbor algorithm (IBK)	Naïve Bayes (NB)	Multilayer perception (ANN)	Proposed hybrid method
First	0.867	0.875	0.813	0.875	0.933
Second	0.667	0.667	0.5	0.667	0.667
ATKT	0.889	1	0.889	1	1
Weighted avg.	0.861	0.91	0.82	0.91	0.941

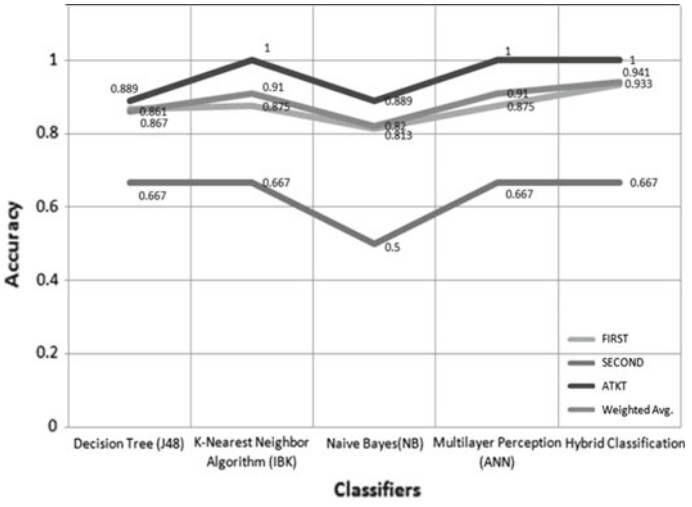

Fig. 5 Comparisons of classifiers accuracy

6 Conclusion

In this paper, the hybrid methodology of classification based on machine learning classifiers to predict performance evaluation is proposed and discussed. The four machine learning algorithms J48, NB, IBK, and ANN were used and ensembled through voting. The result of this paper indicated that the hybrid method of classi-

fication achieved better accuracy as compared to individual machine learning algorithms J48, NB, IBK, and ANN. This hybrid method achieved highest accuracy of 92.59% and individual classifiers, i.e., J48, NB, IBK, and ANN achieved an accuracy of 85.18, 81.48, 88.88, and 88.88%, respectively. Hence, the proposed hybrid method of classification is useful to predict the result of students and it can be used to develop strategies in education to improve performance. The model can also be applied in other domain of data mining and research can be further useful on other ensemble methods for classification of data sets.

References

1. B. Bakhshinategh, O.R. Zaiane, S. ElAtia et al., Educational data mining applications and tasks: a survey of the last 10 years. Educ. Inf. Technol. (2017)
2. C. Silva, J. Fonseca, Educational data mining: a literature review, in *Europe and MENA Cooperation Advances in Information and Communication Technologies*, vol. 520, Advances in Intelligent Systems and Computing, ed. by A. Rocha, M. Serrhini, C. Felgueiras (Springer, Cham, 2017)
3. R. Baker, Data mining for education, in *International Encyclopedia of Education*, vol. 7, 3rd edn., ed. by B. McGaw, P. Peterson, E. Baker (Elsevier, Oxford, UK, 2010), pp. 112–118
4. C. Romero, S. Ventura, Educational data mining: a review of the state-of-the-art. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **40**(6), 601–618 (2010)
5. C. Romero, S. Ventura, in *Data Mining in Education, WIREs Data Mining Knowledge Discovery*, vol. 3 (Wiley, 2013), pp. 12–27. https://doi.org/10.1002/widm.1075
6. B. Behdad, R. Osmar, E. Samira, *Educational Data Mining Applications and Tasks: A Survey of the Last 10 Years* (Education Information Technology, Springer, 2017). https://doi.org/10.1007/s10639-017-9616z
7. C. Laura, A. Angel, Educational data mining and learning analytics: differences, similarities, and time evolution. RUSC **12**(3) (2015). Barcelona. July 2015. https://doi.org/10.7238/rusc.v1213.2515
8. T. Abinash, A. Abhishek, K. Santannu, Document-level sentiment classification using hybrid machine learning approach. J. Knowl. Inf. Syst. **53**, 805–831 (2017). https://doi.org/10.1007/s10115-017-1055-z
9. J.R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Francisco, 1993)
10. K. Chhogyal, A. Nayak (2016) An empirical study of a simple Naïve Bayes classifier based on ranking functions, in *AI 2016: Advances in Artificial Intelligence. AI 2016*, vol. 9992. Lecture Notes in Computer Science, ed. by B. Kang, Q. Bai (Springer, Cham, 2016)
11. D.M. Powers, Evaluation: from precision, recall and F-measure to ROC, in *Informedness, Markedness and Correlation* (2011)