2017 8th International Conference on Information Technology (ICIT)

Students Performance Prediction Using KNN and Naïve Bayesian

Ihsan A. Abu Amra
Faculty of Information Technology
Islamic University of Gaza
Gaza-Palestine
ihsanamra2015@gmail.com

Ashraf Y. A. Maghari
Faculty of Information Technology
Islamic University of Gaza
Gaza-Palestine
amaghari@iugaza.edu.ps

***Abstract*— Data mining techniques is rapidly increasing in the research of educational domains. Educational data mining aims to discover hidden knowledge and patterns about student performance. This paper proposes a student performance prediction model by applying two classification algorithms: KNN and Naïve Bayes on educational data set of secondary schools, collected from the ministry of education in Gaza Strip for 2015 year. The main objective of such classification may help the ministry of education to improve the performance due to early prediction of student performance. Teachers also can take the proper evaluation to improve student learning. The experimental results show that Naïve Bayes is better than KNN by receiving the highest accuracy value of 93.6%.**

Keywords—educational data mining; KNN; Naïve Bayes; classification

I. INTRODUCTION

The student performance prediction is a very important issue for improving the educational process. The students' performance level may be affected by many factors related to father's job, gender, and their average marks in the previous years. The early prediction of student performance may help in improving the educational process[1]. The performance prediction can be obtained by employing data mining techniques on educational data sets. The data classification is the most important technique in data mining research. It depends on categorization (giving a class) of data based on the values of the predicting attributes [2]. Classifiers are represented by different kinds of models. And the variation of algorithms is great in some times for inducing classifiers from data. Some popular classification algorithms are K -Nearest Neighbors classifier, Naïve Bayes, Neural networks, Decision Tree Algorithms (ID3, C4.5, and Random Forest), and Support Vector Machine (SVM).

Several efforts provided prediction of students results based on applying classification algorithms. However, a few efforts applied classification algorithms on educational data sets of secondary schools in Gaza Strip. The major objectives of such classification may help the ministry of education to predict the performance of new students based on similar relationships. This can improve the student's performance before they finish their study.

According to statistics of the ministry of education in Gaza Strip for 2015, the students who attend to get the Secondary General Certificate (SGC) are 33294 but 7809 of them did not succeed. This number represents 27% of the total number. Moreover, it represents around 55% of the total number of students in some schools.

This paper proposes a student performance prediction model using KNN and Naïve Bayes as classification techniques applied on data set for secondary General certificate, collected from the ministry of education in Gaza Strip.

The general knowledge discovery process requires three main steps as we will see in section III.

1. A pre–processing: This phase is essential to achieve data quality. So, data reduction is done by selecting the most important attributes without losing quality[3]. Eight attributes of our data set (500 records) are selected from 14 attributes. A list of attributes after preprocessing is presented in Table 1.

*2. Classification process of data mining***:** Some classification algorithms are applied on rapid miner IDE, to split our educational dataset into training and testing data, as we will see in section III.

*3. Post- Processing***:** the ministry of education in Gaza Strip may use the classification model in order to predict the performance of new students based on similar relationships used in our data set and to improve the student's performance before they finish their study.

This paper is organized as the following: in section II, we explain some related works; section III presents a background, section IV demonstrates the methodology of applying classification techniques on dataset and the results. The last section contains the conclusion and future work.

978-1-5090-6332-1/17/$31.00 ©2017 IEEE

II. RELATED WORK

Applying data mining techniques on educational data can help in identifying, extracting, and evaluating variables related to learning process of students [4], especially when related features are selected to the learning process. Some researchers [4] applied data mining techniques (classification, clustering and outlier detection rules) to extract important knowledge that can be used in the educational domain. They applied .e.g. naïve Bayesian classifier on the dataset and obtained a prediction model with accuracy of 67.50%. They also suggested this classifier for predicting the student grade on time.

Jadhav and Channe [5] conducted two experiments to predict the student's final mark. The first experiment compared classification algorithms using three datasets in which the results achieved with better accuracy when all available data are taken into account versus filtering. In the second experiment, they concluded that the best accuracy can be obtained by applying classification model for both numerical and categorical data.

Some researchers [6][7][8] focused on comparing various classification techniques and reported the advantages and disadvantages for each one. Other researches [3] conducted their comparison study between decision tree and Naïve Bayes algorithms using orange tool to predict bandwidth pattern in different intervals among different users in the network. They used the precision, recall and accuracy as measurements and concluded that the decision tree has the highest performance. Moreover, Ogunde and Ajibade [9] developed a system using Iterative Dichotomiser (ID3) decision tree algorithm to predict students' graduation grades based on entry results data. The authors declared that their system could be very useful in predicting student's final graduation grades.

Bhardwaj and Pal, [10] proposed a research work to find out that any pattern could be useful as the strongest prediction for students performance at the UNWE university based on pre-university and characteristics of persons. The results showed that the decision tree performs the best accuracy, followed by KNN classifier whereas the Bayes classifiers had the lowest accuracy. The authors analyzed the data gathered from colleges and applied classification techniques for student performance prediction by using Matlab tools. Their study showed that the performance of the students does not always depend on the student effort, but, other factors can affect the student's performance.

Cortez and Silva [11], addressed the prediction of secondary student grades of both (mathematics and portugueue language). Their work intended to approach the achievement of student in secondary education using BI/DM techniques. Their obtained results reveal that if the grade of

first and second school are known, a high predictive accuracy can be achieved.

Recently, other researchers [12][13] conducted a study to predict the performance of academic students in both master's and bachelor's degree using different classification algorithms and different subjects.

However, a few efforts applied classification algorithms on educational datasets of secondary schools in Gaza Strip. In our study we applied major kinds of classification techniques KNN and Naïve Bayes to help the ministry of education to predict the performance of new students, and improve results for the next year.

III. BACKGROUND

A. K-Nearest Neighbors classifier(KNN)

The categorization of unknown data point is the basis of nearest neighbor in which its class is already known, this is the simplest definition of KNN.

In this algorithm the nearest neighbor is calculated according to k-value that determines the number of nearest neighbors to be considered and hence defining the class of a sample data point [6]. Sometimes, it depends on the use of more than one nearest neighbors to determine the class of the given data point belongs to that is the reason for calling: KNN. This algorithm is referred to as memory–based technique because data points must be in the memory at the runtime.

Some researchers [6] improved KNN according to their distances from new sample data point. But memory requirement and computational complexity remain the main concern always. When we reduce the size of data set we can overcome the limitation of memory. So we can eliminate repeated patterns from training samples. To further improve the dataset, some data points can be also eliminated from data set, and those data points don't affect the result.

Some algorithms increase the speed of basic KNN algorithm e.g. ball tree, k-d tree, nearest feature line(NFL), tunable metric, orthogonal search tree and principal axis search tree. To more understand of KNN algorithm, suppose that an object is sampled with a set of different attributes, but the group to which the object belongs is unknown. Determining the class of a sample depends on evaluating the k-number of closest neighbors.

KNN nearest neighbor classification algorithm can be expressed as the following pseudo code:

K← the number of nearest neighbors

For each object Z do

Calculate the distance between every object x and z in the training set d(x, z)

Neighborhood←the k neighbors, closest to z in the training set

Zclass←select class (according to neighborhood)End for

Fig. 1.pseudo code of KNN nearest neighbor classification algorithm

KNN is the simplest of all machine learning algorithms. It has got a wide applications in different fields such that, pattern recognition, marketing of internet, analysis of Image databases cluster, etc.

Sometimes it is helpful to avoid tied votes by choosing k to be an odd number. A single number k is given to determine the total number of neighbors used for classification. When k=1, then the nearest neighbors for a sample will determine its class. KNN require an integer k, a training data set and a metric to measure closeness [8].

Advantages and Disadvantages of KNN [5]:
Advantages

- Ease of understanding and implementing.
- Fast training.
- It is strong to noisy training data.
- It is good when a sample of many class labels.

Disadvantages:

- It is Lazy learners
- It is sensitive to the local structure of the data.
- Memory cost.
- It runs slowly, as it is supervised lazy learner.

B. Naïve Bayes Algorithm

The Naïve Bayes classifier technique is based on Bayesian theorem, whereas it performs better when data dimensionality is high [8]. The Bayesian classifier is capable of calculating the most possible output based on the input. There is no problem to add new raw data at run time and have a better probabilistic classifier.

In this algorithm, the presence of a particular feature in a class is unrelated to the presence of any other feature. Let us describe by example why this algorithm is called a naïve. We judge a fruit is an apple when its characteristics are: round 3 inches in diameter and red, even it depends on each other or on other features, all of these properties independently contribute to the probability to judge that fruit is apple [8].

Bayesian theorem provides an equation for calculating posterior probability P(c|x) from P(c), P(x) and P(x|c):

$$p(c \mid x) = \frac{p(x \mid c)p(c)}{P(c)} \qquad (1)$$

- $P(c|x)$: the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$: the prior probability of *class*.
- $P(x|c)$: the likelihood, which is the probability of *predictor* given *class*.
- $P(x)$: the prior probability of *predictor*.

Let's have a training data set of Average. Now, we need to classify if students will travel or not according to average, based on average condition. To apply naïve Bayes follow these steps:

First step: creating a frequency table from the data set, as shown in TABLE1.
Second step: the probability of Very Good = 0.29 and travel probability is 0.64, this is called Creating Likelihood table as shown in Table 1, (finding the probabilities).
*Third step***:** use Naive Bayes equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.
Problem: Is this statement is correct? Students will travel if Average is Excellent,(that's in our example).

We can solve it using above posterior probability, equation (1). P(Yes | Excellent) = P(Excellent | Yes) * P(Yes) / P (Excellent)

P (Excellent |Yes) = 3/9 = 0.33, P(Excellent) = 5/14 = 0.36, P (Yes)= 9/14 = 0.64.Now, P (Yes | Excellent) = 0.33 * 0.64 / 0.36 = 0.60, It has higher probability.

TABLE1: CREATING A FREQUENCY TABLE FROM THE DATA SET.

Data Set	
Average	**Travel**
Excellent	No
Very good	Yes
Good	Yes
Excellent	Yes
Excellent	Yes
Very good	Yes
Good	No
Good	No
Excellent	Yes
Good	Yes
Excellent	No
Very good	Yes
Very good	Yes
Good	No

Frequency Table		
Average	No	Yes
Very good		4
Good	3	2
Excellent	2	3
Grand total	5	9

Advantages: $\alpha + \beta = \chi.$ (1) (1)

- Short computational time for training.
- Removing the irrelevant features will improve the performance of classification.
- Its performance is good.

Disadvantages:

- The Naïve Bayes classifier requires large number of records to get good results.
- May be less accurate than other classifiers applied on some datasets.

IV. EXPERIMENTS AND RESULTS

In our paper we measured the performance of two classification algorithms (K-Nearest Neighbors classifier, Naïve Bayes) using the same data set. We have carried out two experiments on the same IDE in order to evaluate the performance and usefulness of different classification algorithms to predict students results. We applied three steps of general knowledge discovery process: pre-processing, classification process of data mining then post–processing. Our objective is to classify students according to their attributes into either pass (when student is successful) or incomplete (in case of failure).

A. Data pre–processing

Data-processing, is essential to achieve data quality and very important for Naïve Bayes [14] and other classification algorithms. A data set of 500 records has been selected from more than 2000 records with 8 attributes of students records collected from the ministry of education. The related attributes were selected by observation. TABLE shows a summary of selecting the most important attributes for our experiment.

Our data set is for marks of students in Gaza for 2015 year of both male and female as a result of SGC. So, a pre-processing is essential to achieve data quality that contains cleaning, integration, reduction, and transformation as different technique for data pre-processing.

TABLE 2: A SUMMARY TABLE OF SELECTED ATTRIBUTES OF DATA SET

Column name	Description	Selected
Student_ID	ID for each student	
Student Name	Name of student	
Gender	Female/male	✓
DOB	Date of birth	✓
Place_Of_Birth	Place_Of_Birth for	
Specialization	Student branch(e.g. science)	✓
Enrollment year	The year of studying	
City	Location	✓
Telephone number	Number of telephone	
Secondary school name	secondary school name	✓
Status	e.g. married-single	✓
Fathers 'job	0 if father does not work, 1 means s/he works	✓
Student Status	As pass or incomplete	✓

We removed some irrelevant attributes, and then the selected attributes are processed via rapid miner IDE as declared in Table2. The attributes that have importance in prediction of students' averages for next year are selected. These attributes related to e.g. the job of father, the average of student for a previous year, student school, and so on. The data reduction is done by selecting the most important attributes without losing quality[3].

B. Data Classification

To apply some classification algorithms on rapid miner IDE, the data set should splitted into training and testing data. In our study, the data set is divided into 70% training data and 30% testing data, then, the labeled features have been selected to apply KNN and Naïve Bayes. Precision, recall and accuracy of results are measured for the two applied algorithms.

Table 3 shows the results of applying KNN and naïve Bayesian on the dataset. The accuracy, recall, and precision are used as measurements for the performance of two classifiers in our experiment.

In our experiments, precision means the proportion of data which is classified correctly. And recall means the percentage of information relevant to the class and its correct classification. While accuracy is a percentage of instances classified correctly by classifier. Naïve Bayesian algorithm had the highest accuracy of 93.17%, which is better than KNN. This high percentage means that there is a strong relation between the features used in training process, which affects the student's performance as shown in Table 4.

TABLE3: THE RESULTS OF KNN AND NAÏVE BAYES

	Time	*Size of data set*	*Recall*	*Precision*	*Accuracy*
KNN	*0 Second*	*500 instances*	*62.9*	*63.4*	*63.45*
Naïve Bayes	*0 Second*	*500 instances*	*93.6*	*94.65*	*93.17*

The resulted classification models give relatively accurate prediction, so the ministry can achieve progress in education based on similar attributes for next year e.g. fathers job, the city, gender, the average of students for previous year. All of these features are important for student performance. Our results do not mean that Naïve Bayes is better than KNN in case of changing the data set or used tool. Motwani and Kanojia [15] applied KNN and Naïve Bayes classifiers on similar data set to find the optimal result. Their results showed that KNN classification method gives better accuracy (approx 83.65%) when compared to Naïve Bayes classifier which gives the accuracy of approx 75.77%.

C. Post-*Processing*:

The teachers could use the classification model in order to classify new students according to their performance. Further, the teachers may early advice the students to improve their performance. This prediction may also help to identify key acceptance in universities in which the performance of the student results will be determined previously.

V. CONCLUSION AND FUTURE WORK

Educational database contains amount of data is increasing rapidly. To get the knowledge about student performance, the classification algorithms are applied to the educational

datasets. The study focused on KNN and Naïve Bayes as major classification algorithms to propose a student performance prediction model. In our presented study, by comparing the three evaluate parameters(accuracy, Recall and Precision) for two algorithms KNN and Naïve Bayes, Naïve Bayes algorithm had the highest accuracy 93.17% which means a strong relationship between features that affects the performance of students, and will help for prediction of students performance for the next year. Naïve Bayes is better than KNN, which means a strong relation between features that affect the performance of students, and that will help for prediction of student's performance. Sometimes, KNN will be better than Naïve Bayes for other data sets and different IDE[5]. As future work, more classification algorithms can be applied on different educational datasets.

REFERENCES

[1] V. S. Warke and R. S. Kamath "Data Mining Approach for the Analysis of Performance Based Appraisal System of Selected Teachers in Kolhapur City," no. Iv, pp. 1–6, 2016.

[2] P. Kaur, M. Singh, and G. Singh, "Classification and prediction based data mining algorithms to predict slow learners in education sector," Procedia - Procedia Comput. Sci., vol. 57, pp. 500–508, 2015.

[3] F. Haghanikhameneh, P. H. Shariat Panahy, N. Khanahmadliravi, and S. A. Mousavi, "A comparison study between data mining algorithms over classification techniques in Squid dataset," Int. J. Artif. Intell., vol. 9, no. 12 A, pp. 59–66, 2012.

[4] M. M. A. Tair and A. M. El-halees, "Mining Educational Data to Improve Students ' Performance : A Case Study," Int. J. Inf. Commun. Technol. Res., vol. 2, no. 2, pp. 140–146, 2012.

[5] S. D. Jadhav and H. P. Channe, "Comparative Study of K-NN , Naive Bayes and Decision Tree Classification Techniques," vol. 5, no. 1, pp. 2014–2017, 2016.

[6] T. N. Phyu, "Survey of Classification Techniques in Data Mining," Int. MUlticonference Eng. Comput. Sci., vol. I, pp. 18–20, 2009.

[7] S. Sagunthaladevi, B. Raju, and V. Rama, "SURVEY ON CLASSIFICATION," vol. 5, no. 3, pp. 169–172, 2016.

[8] S. S. Nikam, "Acomparative Study of Classification Techniques in Data Mining Algorithms," Orient. J. Comput. Sci. Technol., vol. 8, no. 1, pp. 13–19, 2015.

[9] Ogunde and Ajibade, "A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm Ogunde A. O 1 . and Ajibade D. A 1 .," Comput. Sci. Inf Technol., vol. 2, no. 1, pp. 21–46, 2014.

[10] B. K. Bhardwaj, "Data Mining : A prediction for performance improvement using classification," Int. J. Comput. Sci. Inf. Secur., vol. 9, no. 4, 2011.

[11] P. Cortez and A. Silva, "Using Data Mining To Predict Secondary School Student Performance," 5th Annu. Futur. Bus. Technol. Conf., vol. 2003, no. 2000, pp. 5–12, 2008.

[12] F. Ahmad, N. H. Ismail, and A. A. Aziz, "The prediction of students' academic performance using classification data mining techniques," Appl. Math. Sci., vol. 9, no. 129, pp. 6415–6426, 2015.

[13] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," Procedia Technol., vol. 25, pp. 326–332, 2016.

[14] P. Chandrasekar and K. Qian, "The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier," 2016 IEEE 40th Annu. Comput. Softw. Appl. Conf., vol. 2, pp. 618–619, 2016.

[15] D. Kanojia, "Comparison of Naive Basian and K-NN Classifier," vol. 65, no. 23, pp. 40–45, 2013.