Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor

Huda Al-Shehri, Amani Al-Qarni, Leena Al-Saati, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi, Sunday O. Olatunji

College of Computer Science and Information Technology, University of Dammam, Saudi Arabia

Abstract—This work presented two prediction models for the estimation of student's performance in final examination. The work made use of the popular dataset provided by the University of Minho in Portugal, which relate to the performance in math subject and it consists of 395 data samples. Forecasting the performance of students can be useful in taking early precautions, instant actions, or selecting a student that is fit for a certain task. The need to explore better models to achieve better performance cannot be overemphasized. Most of earlier work on the same dataset used K-Nearest Neighbor algorithm and achieved low results, while Support Vector Machine algorithm was rarely used, which happens to be a very popular and powerful prediction technique. To ensure better comparison, we applied both Support Vector Machine algorithm and K-Nearest Neighbor algorithm on the dataset to predict the student's grade and then compared their accuracy. Empirical studies outcome indicated that Support Vector Machine achieved slightly better results with correlation coefficient of 0.96, while the K-Nearest Neighbor achieved correlation coefficient of 0.95.

Keywords—Student Performance; Support Vector Machine; K-Nearest Neighbor; Regression.

I. INTRODUCTION

Forecasting student performance is essential for educators to obtain early feedback and take immediate action or early precautions if necessary to improve the student's performance. This prediction can be managed by locating the source of the problem. Should it be from extra activities that the student is participating in, family problems, or health problems. All these factors can have a major effect on student performance. By means of having a dataset for student's performance can help us study such cases. The used dataset in this research paper is collected from two Portuguese secondary schools and is specified in the subject of Math. In this research paper, we used two regression algorithms which are K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) to predict the final grade of the student which falls in the range from 0 to 20. Furthermore, we reviewed 10 previous researches, which predicted the student's performance.

S. B. Kotsiantis et al [1] showed that the best algorithms to predict the student's performance are naive bays and ripper, where Naïve Bayes has some advantages over RIPPER. Raheela Asif et al [2] used Decision Tree, k-nearest algorithms, Rule Induction, Naïve Bayes and Artificial Neural Networks in order to predict the student's performance at university based on high school grades, and to predict the courses that mostly effect the performance of students in the first two years of university. Several data mining techniques were used for classification and the best results were obtained by 1-NN and Naïve Bayes. Dorina Kabakchieva [3] used four classifiers which are Decision Tree classifier, Bayesian classifiers, k-NN Classifier, and Rule learners to discover patterns in their dataset that would assist in predicting the student's performance. 100 and 250 were used as KNN values and achieved an accuracy of about 60%. The results show that (J48) decision tree classifier has the highest overall accuracy than JRip rule learner and K-NN. Havan Agrawal and Harshil Mavani [4] trained a Neural Network to detect academic students who are at risk. The study results show that the three attributes which are the students's grade in secondary education, living location, and medium of teaching, were the highest potential variables with probabilities 0.8642, 0.7862 and 0.7225 respectively. Paulo Cortez and Alice Silva [5] tested Decision Trees, Random Forests, Neural Networks, and Support Vector Machines for both classification and regression. Their results show that student's first and second term grades and number of past failure has the most influence. Following these three attributes are the number of absences, mother's job, and alcohol consumption. The root mean squared error results for SVM regression prediction are 2.09, 2.90, and 4.37 for different input settings. In our experiment, the error results were much lower (with a value of 1.22). Abeer Badr El Din Ahmed and Ibrahim Sayed Elaraby [6] used ID3 Decision Tree classification method to predict the student's final grade and distinguish students who are at risk. Brijesh Kumar Baradwaj and Saurabh Pal [7] used ID3 Decision tree to predict the students' performance. Anal Acharya and Devadatta Sinha [8] used C4.5, MLP, NB, 1-NN, SMO machine learning algorithms to predict the student's performance. The results for obtained for 1KNN were 79% for training, and 66% for testing. Behrouz Minaei-bidgoli et al [9] used Quadratic Bayesian classifier, Parzen-window, 1-NN, KNN, Multilayer Perceptron, and Decision Tree to predict the final grade of the student. The results show that KNN achieved the best performance, in the case of 2-classes, with an accuracy of 82.3%. Mrinal Pandey and S. Taruna [10] used four popular ensemble techniques to predict the students's academic performance which are Adaboost, Bagging, Random Forest and Rotation Forest. The paper concludes that the performance of RTF ensemble was the optimum among the algorithms used, in terms of the accuracy of the model and class, TPR, FPR and ROC curves.

978-1-5090-5538-8/17/$31.00 ©2017 IEEE

From the literature review above, we notice SVM regression algorithm is rarely used, while KNN algorithm received more attention. Since we could not see many researches working on both SVM and KNN we decided to use them together to predict the student's grade, then compare their results to know which one of them is better for regression.

We first needed to convert the dataset from nominal to numeric to find the statistical values (mean, median, and standard deviation, maximum and minimum) and correlation coefficient, as it is not available with nominal attributes. Secondly, we performed the attribute selection process to find the attributes that mostly affect the result. Furthermore, we applied K-NN and SVM regression algorithms using different partition ratios and 10-cross validation. Finally, we settled on 19-NN and Manhattan distance method for K-NN, while for SVM we used epsilon-SVR (regression) type and linear kernel. Empirical study results indicated that SVM slightly outperformed K-NN at this type of problem which showed a higher correlation coefficient value of 0.96. The optimum results for SVM were achieved with percentage split of % 90. As for K-NN, the best results were obtained by using cross validation Folds 10.

The remaining part of this work is organized as follow. Section II contains empirical studies that include dataset description and experimental setup. Section III contains the optimization strategy and parameter search strategy. Section IV presents results and discussion, while section V contains the conclusion and recommendation emanating from this work.

II. EMPIRICAL STUDIES

A. Description of the Dataset

The dataset provided [5] by the University of Minho in Portugal, was collected during the school years 2005-2006, using reports and questionnaires from two Portuguese secondary schools, specified in the subject of Math. During the year, students are evaluated in two periods (G1, G2) and the 3rd period (G3) combines both previous periods to acquire the final score. The grading method used in Portuguese education is a 20-point grading scale where zero indicates the lowest grade and 20 is the highest grade. The dataset includes 33 attributes. Four of the attributes are nominal, 13 of the attributes are binary and 16 of the attributes are numeric. The dataset includes 395 instances with no missing values.

B. Experimental Setup

The experiment is carried out using *Weka* –An open source machine learning software–[11] using *LibSVM* for Support Vector Machine and *IBK lazy classifier* for K-Nearest Neighbor. Moreover, to conduct the statistical analysis, we converted the nominal and binary attributes to numerical attributes, as shown in Table 1, and calculated the mean, median, standard deviation, maximum, and minimum values of the dataset. The percentage of the correlation coefficient of 3-NN was as low as 0.366 and improved to 0.42 after converting the attributes from nominal to numeric. Furthermore, we identify important features that contribute the most to their educational success using correlation based feature selection.

Table 1: Attribute Conversion

Attribute	Description	Conversion
School	Gabriel Pereira or Mousinho da Silveira	GP = 1, MS =2
Sex	Student's gender	F = 1, M = 2
Address	Urban or Rural	U = 1, R =2
Famsize	Family Size	GT3 = 1, LE3 = 2
Pstatus	Parent's cohabitation status (Living together or apart)	A = 1, T = 2
Mjob	Mother's job	at_home = 1, health = 2, other =3, services = 4 , teacher =5
Fjob	Father's job	at_home = 1, health = 2, other =3, services = 4 , teacher =5
Reason	Reason to choose this school	course = 1, other =3, home =2 , reputation =4
Guardian	Mother, father, or other	mother =1 , father =2, other =3
Schoolsup	Extra educational support	yes =1, n =2
Famsup	Family educational support	yes =1, n =2
Paid	Extra paid classes	yes =1, n =2
Activities	Extra-curricular activates	yes =1, n =2
Nursery	Attended nursery school	yes =1, n =2
Higher	Peruses a higher education path	yes =1, n =2
Internet	Internet access at home	yes =1, n =2
Romantic	In a romantic relationship	yes =1, n =2

III. OPTIMIZATION STRATEGY

In this section we start by making changes on each parameter separately and notice the effects each parameter contributes.

A. K-Nearest Neighbor

K-NN technique has seven parameters in Weka: *K-NN, cross Validate, debug, mean Squared, nearest Neighbor Search Algorithm, and Window Size*. We experimented with different values of K to achieve the best correlation coefficient as shown in Table 2. In addition, we changed *Nearest Neighbor Search Algorithm* parameter from the default *Euclidean Distance* function to *Manhattan function* which improves the correlation coefficient. Moreover, the accuracy decreases after increasing the WindowSize to 1. The highest correlation is achieved when

K-NN value is increased to 19, with a correlation value of 0.612. Furthermore, we notice 19KNN achieves the maximum level of accuracy when Manhattan distance function is applied, with the correlation coefficient of 0.67 and relative absolute error of 74.73%. Therefore, to achieve the best performance we gathered the optimum parameters (19 KNN and *Manhattan distance* function) as shown in Table 3, which gave the best accuracy with a decrease in the relative error and an increase in the correlation coefficient

Table 2: Parameter values achieved using different values of K

	Correlation coefficient	**Mean Absolute error**	**Root mean squared error**	**Relative absolute error**	**Root absolute squared error**
3 KNN	0.418	3.195	4.27	92.85%	93.02%
5 KNN	0.512	2.97	3.93	86.33%	85.68%
7 KNN	0.563	2.85	3.82	82.83%	83.14%
9 KNN	0.576	2.825	3.80	82.09%	82.83%
13 KNN	0.579	2.839	3.85	82.49%	83.87%
19 KNN	0.612	2.82	3.845	82.07 %	83.73%

Table 3: Optimum parameters for K-NN model

Parameters	**Optimal values chosen**
KNN	19
cross Validate	False
Debug	False
distance Weighting	No distance weighting
mean Squared	False
Nearest Neighbor Search Algorithm	Linear NN Search (Manhattan distance function)
Window Size	0

B. Support Vector Machine

We optimized parameters for SVM which include: *cost*, *degree*, *eps*, *gamma*, *kernel Type*, and *loss*. The correlation coefficient with the default parameters (Only changing the SVM Type to epsilon-SVR type since we are working with numerical values) is 0.86 with absolute error of 48.89%. As previously, we experiment each parameter separately. We notice the improvement occurs only when *Kernel Type* is changed to *linear: u'*v*, which achieved the maximum level of accuracy with correlation coefficient of 0.9 and relative absolute error of 29.5%. On the other hand, the accuracy decreases after increasing *gamma* to 1. As for the rest of the parameters, no change in accuracy was shown. As a result, we selected the *linear kernel Type* parameter seeing that it gives the highest accuracy, while keeping the rest of the parameters in their default values as shown in Table 4. Comparing the new result with the result of the default parameters shows a significant decrease in relative error and an increase in correlation coefficient.

Table 4: Optimum parameters for the SVM model

parameters	**Optimal values chosen**
SVM type	Epsilon SVR
cost	1
degree	3
eps	0.001
gamma	0
kernel Type	linear: u'*v
loss	0.1

IV. RESULTS AND DISCUSSION

In this section, we discuss the results of feature selection, direct partitioning, and 10-fold cross validation.

A. Results of Investigating the Effect of Feature Selection on the Dataset

The feature selection carried out was the correlation based feature selection. It was attained by calculating the correlation between each attribute and the target variable (G3: final grade). Features selected are based on the values of their correlation with the target variable. SVM and K-NN classifiers have been tested on the dataset using different combinations of feature subsets as illustrated in Table 5. As a result, the feature selection of different features produces exact results for SVM and close results for KNN.

Table 5: Results of different features subset

Number of features	**Features**	**19-KNN**	**SVM**
16 features	G2, Fedu, Mjob, G1, Gout, Medu, FreeTime, Reason, Fjob, StudyTime, Walc, Farmel, Health, Guardian, TravelTime, Dalc	0.77	0.906
8 features	G2, Fedu, Mjob, G1, Gout, Medu, FreeTime, reason	0.81	0.905
4 features	G2, Fedu, Mjob, G1	0.85	0.905
2 features	G2, Fedu	0.87	0.905

B. Results of Investigating the Effect of Different Direct Partition ratio on the performance of the proposed Techniques

After obtaining the results from the feature selection process, which are the second term grades and father's education, we applied K-NN and SVM classifiers using different partition ratios (training: testing). We found that the highest correlation

coefficient in both classifiers is obtained with 90:10 ratio (90% of the data used for training and 10% of the data used for testing) as shown in Table 6.

Table 6: Results of different direct partition ratio

Partition ratio	KNN	SVM
50:50	0.88	0.89
60:40	0.88	0.89
70:30	0.91	0.92
80:20	0.91	0.91
90:10	0.95	0.96

C. Results of Comparing 10-Fold cross validation with Direct Partition methods

From the experimental results above, we select the optimum options from the best features, and partition ratio or cross validation to develop the final model as shown in Table 7.

Table 7: Results of 10-fold cross validation and direct partition ratio

Technique	KNN	SVM
10-Fold Cross Validation	0.87	0.905
Direct Partition in the ratio	0.95	0.96

D. Further Discussions

Previously, we developed the final model using the optimal parameters obtained. The ideal results were achieved with percentage split of 90:10 for SVM and K-NN. The final results showed that SVM slightly exceeds K-NN for this type of problem due to the resulted higher correlation coefficient value of 0.96 and relative absolute error of 19.92%.

V. CONCLUTION AND RECOMMENDATION

Throughout the experiment, we have implemented SVM regression and IBK classifiers on the student dataset to predict the grade of the student ranging from 0 to 20. We then compared the results of both classifiers which were of a high correlation coefficient of 0.96 and 0.95 respectively. Based on the results, we can conclude that both SVM and KNN regression would suit this type of problem. Furthermore, previously published papers were mostly based on predicting the grade based on three values (Bad, Good, Excellent) and the results varied between 52% to 82.3% of accuracy. Even though our work predicted actual values which is regression problem that is more complicated compared to classification problem, yet we were still able to achieve better results. Therefore, looking at the results we got in this case, we recommend to other researchers to look into the techniques used here with the hope of using the methods to solve other problems that could be applicable. Efforts could be made in the near future to also investigate the performance of the proposed techniques on the classification aspect of this problem of performance predictions.

REFERENCES

[1] S. B. Kotsiantis, C. J. Pierrakeas, I. D. Zaharakis, and P. E. Pintelas, "Efficiency Of Machine Learning Techniques In Predicting Students' Performance In Distance Learning Systems," *Educ. Softw. Dev. Lab. Dep. Math. Univ. Patras, Greece*, pp. 297–305, 2003.

[2] R. Asif, A. Merceron, and M. K. Pathan, "Predicting Student Academic Performance at Degree Level : A Case Study," *I.J. Intell. Syst. Appl.*, pp. 49–61, 2015.

[3] D. Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification," *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 61–72, 2013.

[4] H. Agrawal and H. Mavani, "Student Performance Prediction using Machine Learning," *Int. J. Eng. Res. Technol.*, vol. 4, no. 3, pp. 111–113, 2015.

[5] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[6] A. B. E. D. Ahmed and I. S. Elaraby, "Data Mining : A prediction for Student's Performance Using Classification Method," *World J. Comput. Appl. Technol.*, vol. 2, no. 2, pp. 43–47, 2014.

[7] B. K. Baradwaj and S. Pal, "Mining Educational Data to Analyze Students' Performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 6, pp. 63–69, 2011.

[8] A. Acharya and D. Sinha, "Early Prediction of Students Performance using Machine Learning Techniques," *Int. J. Comput. Appl. (0975 – 8887)*, vol. 107, no. 1, pp. 37–43, 2014.

[9] B. Minaei-bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "Predicting Student Performance : An Application Of Data Mining Methods With The Educational Web-Based System," *33rd ASEE/IEEE Front. Educ. Conf.*, vol. 1, pp. 1–6, 2003.

[10] M. Pandey and S. Taruna, "A Comparative Study of Ensemble Methods for Students' Performance Modeling," *Int. J. Comput. Appl. (0975 – 8887)*, vol. 103, no. 8, pp. 26–32, 2014

[11] WEKA. Hamilton, New Zealand, 1999