Turkish Journal of Science & Technology
Volume 13(2), 53-59, 2018

Data Mining Techniques Based Students Achievements Analysis

Dönüş Şengür[1], Songül Karabatak[2]
[1]Milli Eğitim Bakanlığı, Şht.P. Bnb.Zafer KILIÇ Anadolu İmam Hatip Lisesi, Elazığ
[2]Fırat Üniversitesi, Enformatik Bölümü, Elazığ
[*]kdksengur@gmail.com

(Geliş/Received: 13.07.2018;Kabul/Accepted: 07.08.2018)

Abstract

In this work, data mining techniques are used to determine the students' achievements in Mathematic class. In other words, we use the data mining techniques to determine if there is any link between the student achievement and various student related data such as student grades, demographic, social information and school related data. Data mining techniques, Decision Tree (DT), Discriminant Analysis (DA), Support Vector Machines (SVM), k-nearest neighbor (K-NN) and ensemble learner are used in prediction purposes. A publicly available dataset is considered in experimental works. Experimental works, on computer environment are carried out to validate the data mining techniques. All data mining methodologies are simulated on MATLAB environment with 5-fold cross-validation technique. The classification performance is measured by accuracy and root mean square error (RMSE) criterions. Three experimental setups and for each setup, three scenarios are considered during experimentation. The obtained results are encouraging and the comparison with some of the existing achievements shows the superiority of our work.

Keywords: Data Mining Techniques, Student's Achievements, Classification, Regression.

Veri Madenciliği Teknikleri ile Öğrencilerin Başarım Analizi

Özet

Bu çalışmada, öğrencilerin matematik dersindeki başarımlarının belirlenmesi için veri madenciliği teknikleri kullanılmıştır. Diğer bir ifade ile öğrenci başarısı ile öğrenci notları, demografik, sosyal bilgiler ve okulla ilgili veriler gibi çeşitli öğrenci verileri arasında herhangi bir bağlantı olup olmadığını belirlemek için veri madenciliği teknikleri kullanılmıştır. Karar Ağaçları (KA), Diskriminant Analiz (DA), Destek Vektör Makineler (DVM), k-en yakın komşular (k-EYK) ve birleştirilmiş öğreniciler yöntemleri tahmin amacıyla kullanılmıştır. Deneysel çalışmalarda halka açık bir veri seti kullanılmıştır. Veri madenciliği tekniklerinin doğruluklarını belirlemek için bilgisayar ortamında çeşitli deneysel çalışmalar gerçekleştirilmiştir. Tüm veri madenciliği yöntemleri MATLAB ortamında 5 kat çapraz doğrulama tekniği ile simüle edilmiştir. Sınıflandırma performansı, doğruluk ve ortalama karesel hatasının karekökü (OKHK) kriterleri ile ölçülmüştür. Üç farklı deneysel çalışma ve her bir deney için ise üç farklı senaryo test edilmiştir. Elde edilen sonuçlar cesaret vericidir ve literatürde mevcut olan bazı sonuçlar ile karşılaştırmalar gerçekleştirilen çalışmanın üstünlükleri gösterilmiştir.

Anahtar Kelimeler: Veri Madenciliği Teknikleri, Öğrenci Başarımı, Sınıflama, Regresyon

1. Introduction

Recently, educational data mining (EDM) has become an emerging topic, which attracts a great amount of the researchers [1-10]. Prediction and/or modelling of various quantities related with education can help to develop better educational environments for students. Especially, student's performance prediction based on various quantities such as student grades, demographic, social and school related features, has been quite hot in the last decade with numerous publications. For example; Cortez et al. used several data mining techniques to predict the student achievement in secondary education [1]. Authors used four supervised data mining techniques namely, decision trees (DT), random forests (RF), neural networks (NN) and support vector machines (SVM) and demographic, social and school related features

to predict the achievements of the students. Ma et al. used data mining technique to performance evaluation of the students of Singapore [2]. Association rules (AR) method was used to select weak tertiary students for remedial classes. Demographic and school performance quantities were used as input to the AR method. Minaei-Bidgoli et al. modelled the student grades by data mining techniques [3]. The work was applied on 227 students from Michigan State University and three different data mining approaches were considered. Authors indicated that an ensemble classifier yielded the best result. Sengur et al. used various data mining techniques for prediction of the graduate scores of the university students [4]. The first and second years lecture scores were used to predict the graduate points of the students from education faculty. Authors used DT, NN and k-NN classifiers. Kotsiantis et al. used various data mining methods in order to predict the performance of computer science students from a university distance learning program [5]. Demographic and performance attributes were considered as inputs to the classifiers. Authors mentioned that the best accuracy was obtained with Naive Bayes (NB) method. Pardos et al. used regression methods for prediction of the math test scores of students [6]. Authors collected data from an online tutoring system. The authors indicated that the Bayesian Networks produced the best results. Turhan et al. used data mining techniques for modelling of the school administrator's conflict handling styles [7]. Authors proposed neutrosophic weighted SVM technique and showed its efficiency on educational dataset. Kabakchieva et al. used data mining algorithms for prediction of student performance [8]. Author considered four data mining techniques such as rule learner, DT, NN, and a K-NN. Author also mentioned that the best accuracy was obtained with NN classifier. Devasia et al. also used various data mining techniques for student's performance prediction [9]. Authors used a web based system that uses the data mining techniques for feature extraction. 700 students were used in the experiments and 19 attributes were saved. It was mentioned that NB method yielded better results than DT, NN and regression based methods. Vyas et al. used CART method for student's performance prediction [10]. More specifically, authors used DT approach on student's current performance and some measurable past attributes to detect good and bad performers.

In this paper, various data mining techniques are used for student's performance prediction. The dataset, which was presented in [1], is also used in our work. The dataset attributes, which was collected by using school reports and questionnaires, include student grades, demographic, social and school related features. There are totally 33 attributes where G3 was used as target attribute. Two datasets were provided regarding the performance in two classes: Mathematics and Portuguese language. In [1], the two datasets were modeled under binary/five-level classification and regression tasks that we also follow in our work. Different from in [1], only 10 attributes (famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2 and G3) are considered in our work. In other words, the attributes that have numerical data are considered for modelling. Data mining techniques, DT, discriminant analysis (DA), SVM, K-NN and ensemble learner are used in prediction purposes. Various experimental works, on computer environment are carried out to validate the proposed idea. All data mining methodologies are simulated on MATLAB environment with 5-fold cross-validation technique. The classification performance is measured by accuracy and root mean square error (RMSE) criteria.

The experimental results show that for all of the three modelling tasks (binary/five-level classification and regression), the selected attributes obtain better accuracy and RMSE scores than the result presented in [1] for Mathematics class.

2. Materials and Methods

2.1. Dataset

The dataset, which is used in this work, was produced by Cortez et al [1]. The students who were in two public schools from Portugal during the 2005- 2006 school year were considered for data collection. Questionnaires, paper sheets, three period grades and number of school absences information were used to construct the

attributes of the data set. In addition, demographic, social and school related information were also used in dataset construction. Thus, the dataset contains totally 33 attributes where G3 (final grade) is assigned as target attribute. Two classes namely Mathematics with 395 examples and the Portuguese language with 649 records were collected. In our work, we only use 10 of 33 attributes namely; quality of family relationships (famrel), free-time after school (free-time), going out with friends (gout), workday alcohol consumption (Dalc), weekend alcohol consumption (Walc), current health status (health), number of school absences (absences), first period grade (G1), second period grade (G2) and final grade (G3).

2.2. Data mining techniques

Data mining, which covers a dozens of techniques, aims to extract important knowledge from a given of dataset. This can be done by either supervised or unsupervised way. In supervised classification, a set of labelled dataset is used to figure out some rules that link the input dataset to labels. Thus, the obtained rules can be further used for some unknown dataset to determine their class labels. Data classification and regression can be seen in supervised data mining. DT, discriminant analysis, SVM, k-NN and ensemble learner methods are used in our work. DT is a popular, non-parametric supervised learning method that efficiently classifies datasets [11]. The goal of DT is to create a model that predicts the class label of a test sample by learning simple decision rules inferred from the input data set. Discriminant analysis produces a series of equations based on input feature space that are used to classify a test sample [12]. SVM is another important and efficient supervised classification algorithm [13]. SVM models a decision boundary between classes of training data as a separating hyperplane. In k-NN classification procedure, all training samples are used to classify the test sample according to a pre-defined distance function and the number of nearest neighbor's k [14]. An ensemble learner is known to be constructed from a serious of individual classifiers [15]. In other words, an ensemble learner determines a test sample class label by combining the decisions of the individual classifiers in some ways.

3. Experimental Works and Results

The aim of this work is to predict the student's performance based on several key factors to determine what affects their educational success/failure. To this end a dataset is used that covers Mathematics class. The data mining techniques are used to model the relationship between factors and final grades. Three different scenarios are considered as it was done in [1]. In the first scenario, the binary classification is considered. In the binary classification, the class labels are pass ($G3 \geq 10$) or fail. In the second scenario, the classification is handled with five levels which start with good to insufficient. In the third one, the regression operation is considered with numeric output that ranges between 0 and 20. In addition, three experimental setups (A, B and C) are considered as suggested in [1]. In A, G3 attribute is used as target and the rest attributes are used as input. In B, G3 is also used as target and except G2, all attributes are used for input. C is similar to B where except G1, all attributes are used as input.

The MATLAB software is used in modelling of the data mining techniques. More specifically, the classification learner application is preferred which enables the user to explore supervised machine learning using various classifiers. In the evaluation of the employed data mining techniques, 5-fold cross validation test is used and the mean accuracy values are recorded. For regression analysis, the regression learner application is used and the RMSE value is calculated for performance evaluation. The DTs, which are used in the prediction, are coarse tree, medium tree and fine tree, respectively. DTs have several positive properties such as easy to interpret, fast for fitting and low memory usage. But they can get low predictive accuracies based on the application. In coarse tree structure, the maximum number of splits is assigned as 4. Similarly, the maximum numbers of splits for medium and fine trees are 20 and 100, respectively. Discriminant analysis is a fast, accurate, easy to interpret and effective classification algorithm where linear and

quadratic discriminants analysis are two main forms of it. While linear discriminant creates linear boundaries between predicted classes, quadratic discriminant constructs non-linear boundaries between predicted classes. SVM algorithm seeks the best hyperplane where the separation of data points of one class from others is guaranteed. The classification learner application presents three different SVM algorithms such as linear, quadratic and cubic, respectively. As k-NN classifiers generally have high predictive accuracy in low dimensions. This can be seen as an advantage of the k-NN classifiers, high memory usage, and not easy to interpret properties make them disadvantageous. Three different k-NN techniques are presented in classification learner application such as fine k-NN, cubic k-NN and weighted k-NN, respectively. In fine k-NN technique, the number of neighbor k is chosen as 1. In addition, for medium and coarse k-NN techniques, the numbers of neighbors are selected as 10 and 100, respectively. In cubic k-NN approach, the cubic distance metrics is used and the numbers of the neighbors are set to 10. In weighted k-NN, the number of neighbor is also set to 10 and a weighted distance function is used for class separation. As ensemble learner covers a dozen of classifiers, they are generally announced as slow approaches. Two different ensemble learner techniques are used in classification learner application such as boosted trees and bagged trees, respectively. In boosted trees technique, the AdaBoost ensemble method is used. In addition, for bagged trees, the random forest approach is adopted.

Table 1 shows the obtained results for binary classification setup. The results for A, B and C scenarios are given in the last three columns of Table 1. For scenario A, the best accuracy score 91.6% is obtained with Quadratic discriminant analysis. The second-best accuracy score 91.4% is produced by linear SVM technique. Linear discriminant analysis produces 90.6% accuracy score that is the best third result for scenario A. DT, DA, SVM and ensemble learners methods obtain better results than k-NN methods. The worst accuracy 81.0% is produced by fine k-NN method and cubic and weighted k-NN methods produce the 83.8% accuracy value. Boosted and bagged trees methods produce 90.1% and 89.1% scores, respectively.

Table 1. Binary classification results for Mathematics class. The bold case shows the highest accuracy.

Data Mining Technique	Classifier Type	Accuracy (%)		
		A	B	C
DT	Fine Tree	90.1	82.5	88.4
	Medium Tree	89.9	82.0	88.4
	Coarse Tree	90.4	83.0	**91.4**
Discriminant analysis	Linear Discriminant	90.6	85.3	88.6
	Quadratic Discriminant	**91.6**	81.0	87.6
SVM	Linear SVM	91.4	**86.1**	89.6
	Quadratic SVM	86.6	80.8	86.6
	Cubic SVM	86.1	78.2	84.8
k-NN	Fine k-NN	81.0	72.9	78.0
	Cubic k-NN	83.8	79.0	81.5
	Weighted k-NN	83.8	77.0	82.8
Ensemble learners	Boosted Trees	90.1	82.5	87.8
	Bagged Trees	89.1	83.0	90.6

For the scenario B, the best accuracy is obtained by linear SVM method. The accuracy of linear SVM is 86.1%. The second-best accuracy 85.3% for scenario B is produced by linear discriminant analysis. The fine, medium and coarse tree methods produce 82.5%, 82.0% and 83.0% accuracy scores, respectively. The boosted and bagged trees obtain 82.5% and 83.0% scores, respectively. The worst accuracy scores are produced by k-NN methods. While the fine k-NN method produces 72.9% accuracy score, cubic and weighted k-NN methods obtain 79.0% and 77.0% scores. Cubic SVM technique is also produced one the worst result where the accuracy is 78.2%. For scenario C, the best accuracy 91.4% is obtained with coarse tree method. Other DT methods (fine and medium tree) produce 88.4% accuracy scores. The bagged trees method produces the second-best accuracy score where the calculated accuracy is 90.6% and linear SVM technique obtain the third best achievement. With 78.0% accuracy score, the fine k-NN approach yields the worst accuracy score for scenario C.

Table 2. Five-level classification results for Mathematics class. The bold case shows the highest accuracy.

Data Mining Technique	Classifier Type	Accuracy (%)		
		A	B	C
DT	Fine Tree	73.2	53.2	72.4
	Medium Tree	76.5	53.9	74.2
	Coarse Tree	70.4	56.7	71.4
Discriminant analysis	Linear Discriminant	71.4	58.0	71.6
	Quadratic Discriminant	73.4	57.2	71.9
SVM	Linear SVM	74.9	58.2	**76.5**
	Quadratic SVM	73.4	55.4	71.9
	Cubic SVM	68.6	50.4	65.8
k-NN	Fine k-NN	57.2	45.1	50.1
	Cubic k-NN	55.4	47.8	49.1
	Weighted k-NN	57.5	48.6	53.4
Ensemble learners	Boosted Trees	**78.7**	**61.5**	74.7
	Bagged Trees	73.4	53.7	73.4

The prediction achievements for five-level classification setup are given in Table 2. As seen in Table 2, for scenario A, the best achievement is carried out with boosted trees method, where the produced accuracy is 78.7%. With 76.5% accuracy score, the medium tree method obtains the second-best accuracy for scenario A. Quadratic discriminant, quadratic SVM and bagged trees methods produce the 73.4% third-best accuracy score for scenario A. With 55.4% accuracy score, the cubic k-NN obtains the worst achievement. Boosted trees method is also produced the best accuracy score for scenario B. 61.5% accuracy score is recorded by the boosted trees method. Linear SVM method produces the 58.2% accuracy score for scenario B which is the second-best accuracy. In addition, the linear discriminant method produces the 58.0% accuracy score which is the third-best achievement. The worst result 45.1% is produced by fine k-NN method. Linear SVM method obtains 76.5% accuracy score, which is the best achievement for scenario C of five-level classification setup. Boosted trees method produce the 74.7% accuracy value, that is the second-best achievement and medium tree obtains the 74.2% accuracy score which the best-third achievement. Cubic k-NN produces the worst result where its achievement is 49.1%.

Table 3. Regression results for Mathematics class. The bold case shows the highest accuracy.

Data Mining Technique	Classifier Type	RMSE		
		A	B	C
DT	Fine Tree	1.90	2.49	2.03
	Medium Tree	1.75	2.62	1.86
	Coarse Tree	2.16	2.82	2.11
Linear Regression	Linear	2.10	2.75	1.93
	Robust linear	2.04	2.98	2.04
SVM	Linear SVM	2.00	2.86	2.01
	Quadratic SVM	2.00	2.86	2.01
	Cubic SVM	2.08	3.01	2.03
Ensemble learners	Boosted Trees	**1.72**	**2.42**	**1.76**
	Bagged Trees	1.85	2.62	1.93

The regression results are shown in Table 3 for all A, B and C scenarios. For regression experiments, the DT, linear regression, SVM and ensemble learners techniques are considered and the obtained results were evaluated based on RMSE scores. The lower RMSE score shows the better achievement. As seen in the Table 3, boosted trees method achieves the lowest RMSE scores for all A, B and C scenarios where the RMSE values are 1.72, 2.42 and 1.76, respectively. Medium tree method achieves the second lowest RMSE values where the obtained RMSE values are 1.75, 2.62 and 1.86, respectively and the bagged trees method produces the third lowest RMSE score for A, B and C scenarios, respectively. The bagged trees achievements are 1.85, 2.62 and 1.93, respectively.

We further compare the obtained achievements with the results in [1]. The comparisons are given in Table 4. The first row of Table 4 shows the achievements in [1] and the second row shows our achievements. As seen in Table 4, for only scenario A of the binary classification setup, the Cortez et al.'s achievement is better than ours and in the rest setups and scenarios, our achievements are better than Cortez et al.'s achievements.

Table 4. Comparisons of the results for Mathematics class. The bold case shows the best scores.

Method	Binary classification Accuracy (%)			Five-level classification Accuracy (%)			Regression RMSE		
	A	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Cortez et al. [1]	**91.9**	83.8	70.6	78.5	60.5	33.5	1.75	2.46	3.90
Our Work	91.6	**86.1**	**91.4**	**78.7**	**61.5**	**76.5**	**1.72**	**2.42**	**1.76**

4. Conclusions

The main aim of this work is to determine the key variables, which affect the student's educational achievement (success/failure). To this end, the data mining techniques are used. Some of the attributes from the dataset which was produced in [1] is used in our experimental works. The used attributes are famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2 and G3, respectively. As it was stressed in [1], three different experimental setup and for each experimental setup, 3 various scenarios are considered in our works and the obtained results are evaluated accordingly. 13 data mining techniques are run for binary and five-level classification setups and 10 data mining techniques are used for regression setup. DT, DA, SVM and ensemble learner methods generally outperform and k-NN methods generally tend to produce the worst results. We also compare our achievements with the results in [1]. The comparisons show that our achievements are generally better than the achievements in [1].

5. References

1. Cortez, P., and Silva, A. (2008). Using data mining to predict secondary school student performance. In the *Proceedings of 5th Annual Future Business Technology Conference*, Porto, Portugal, 5-12.
2. Ma, Y., Liu, B., Wong, C., Yu, P., and Lee, S. (2000). Targeting the right students using data mining. In *Proc. of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Boston, USA, 457–464.
3. Minaei-Bidgoli, B., Kashy, D., Kortemeyer, G., and Punch, W. (2003). Predicting student Performance: An application of Data Mining Methods with an educational web-based system. In *Proc. of IEEE Frontiers in Education*. Colorado, USA, 13–18
4. Şengür, D., and Tekin, A. (2013). Öğrencilerin mezuniyet notlarinin veri madenciliği metotlari ile tahmini. *Bilişim Teknolojileri Dergisi*, **6**(3): 7-16.
5. Kotsiantis S.., Pierrakeas, C., and Pintelas P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence (AAI)*, **18**(5): 411–426.
6. Pardos Z., Heffernan N., Anderson B., and Heffernan C. (2006). Using fine-grained skill models to fit student performance with bayesian networks. In *Proc. of 8th Int. Conf. on Intelligent Tutoring Systems*. Taiwan.
7. Turhan, M., Şengür, D., Karabatak, S., Guo, Y., and Smarandache, F. (2018). Neutrosophic weighted support vector machines for the determination of school administrators who attended an action learning course based on their conflict-handling styles. *Symmetry*, **10**(5): 176.
8. Kabakchieva, D. (2012). Student performance prediction by using data mining classification algorithms. *IJCSMR*, **1**(4): 686-690.
9. Devasia, T., Vinushree, T. P., and Hegde, V. (2016, March). Prediction of students performance using educational data mining. In *Data Mining and Advanced Computing (SAPIENCE), International Conference on* (pp. 91-95). IEEE.
10. Vyas, M. S., & Gulwani, R. (2017, April). Predicting student's performance using cart approach in data science. In *Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of* (Vol. 1, pp. 58-61). IEEE.
11. Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology, *IEEE Transactions on Systems Man and Cybernetics*, **21**: 660-674.
12. Sengur, A. (2008). An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases. *Expert Systems with Applications*, **35**: 214-222.
13. Hastie T., Tibshirani R., and Friedman J. (2001). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Springer-Verlag, NY, USA.

14. Fix, E., and Hodges, J. L. (1951). Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties; Technique Report No. 4; U.S. Air Force School of Aviation Medicine, Randolf Field Texas: Universal City, TX, USA, pp. 238–247.

15. Sengur, A. (2012). Support vector machine ensembles for intelligent diagnosis of valvular heart disease. *Journal of Medical Systems*, **36**: 2649-2655.