Soumadip Ghosh, Arnab Hazra,
Payel Biswas, Amitava Nag

OCENA UCZNIOWSKICH WYNIKÓW PRZY WYKORZYSTANIU SZTUCZNEJ SIECI NEURONOWEJ

[**słowa kluczowe:** *data mining*, klasyfikacja, wielowarstwowy perceptron, drzewa decyzyjne, maszyna wektorów wsparcia]

Streszczenie

Ilość danych generowanych co roku w placówkach oświatowych jest ogromna i właśnie ze względu na tę dużą ilość istnieje potrzeba zapewnienia wydajnego wsparcia systemowego, aby ułatwiać podejmowanie właściwych decyzji. Niniejsze badanie dotyczy oceny wyników uczniów z wykorzystaniem techniki *data mining* w danej przestrzeni danych. Baza danych, z której korzystamy w naszym badaniu, jest powiązana z wykształceniem średnim w dwóch portugalskich szkołach. Dostępne były dwa zestawy danych dotyczące oceny efektywności nauczania dotyczących dwu przedmiotów: matematyki i języka portugalskiego. Celem klasyfikacji jest ewaluacja końcowej oceny uczniów w dowolnym instytucie. W naszej pracy korzystamy z wielowarstwowego perceptronu (MLP), będącego symulowanym sztucznym modelem sieci neuronowej, który mapuje zbiory danych wejściowych na zbiór odpowiednich wyników. Pozostałe techniki klasyfikacji używane w tym zbiorze danych to drzewo decyzyjne (DT) i maszyna wektorów wsparcia (SVM). Wyniki pokazują, że wydajność MLP jest lepsza w porównaniu z dwiema pozostałymi technikami.

* * *

Soumadip Ghosh, Arnab Hazra, Payel Biswas, Amitava Nag

STUDENTS' PERFORMANCE EVALUATION USING ARTIFICIAL NEURAL NETWORK

[**keywords**: Data Mining, Classification, Multilayer Perceptron, Decision Tree, Support Vector Machine]

Abstract

The volume of data generated every year in educational institutions is enormous; due to this large volume of data there is a need to provide an efficient system support to aid in good decision making process. This research study is all about the evaluation of student performance using data mining technique over a given data space. The database that we are using in our study is related with the secondary education of two Portuguese schools. Two datasets are provided regarding the performance evaluation in two distinct subjects: Mathematics and Portuguese language. The classification goal is to evaluate the final grade of the students in any institute. In our work we are using Multilayer Perceptron (MLP), which is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. The other classification techniques that are used over this dataset are Decision Tree (DT) and Support Vector Machine (SVM). The performance of MLP is found to be superior compared to the other two techniques used here.

1. Introduction

There are increasing research interests in using data mining [1] in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data come from educational environments. One way to achieve highest level of quality in higher education system is by discovering knowledge for prediction regarding enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students' performance and so on. The knowledge is hidden among the educational data set and it is extractable through data mining techniques**.** This work is all about the evaluation of student performance using data mining technique over a given data space. The database that we are using in our work is related with secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics and Portuguese language.

Our classification [2] goal is to evaluate the final grade of the students in any institute. In our project we are using Multilayer Perceptron (MLP)[3], which is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique technique called back propagation for training the network.

The other two classification techniques that we will be using to compare the results are Decision Tree and Support Vector Machine (SVM).

Decision tree [4] [5] is a classification model in which a decision tree learns from the tuples in the training dataset. A decision tree appears like a flowchart in a tree like structure, where each internal node denotes condition testing on an attribute, each branch resulting from that node denotes the outcome from the test. The leaf node in the decision tree holds a class label. In this tree, the nodes divide the tuples into different groups at each level of the tree until they fall into distinct class labels.

Whereas Support vector Machine (SVM) [6] uses nonlinear mapping to transform the linear dataset into a higher dimension. In this dimension it searches for the linear optimal separating hyperplane. A hyperplane is a decision boundary to separate two classes. Support vectors are the essential training tuples from the set of training dataset. With a sufficiently high dimension and appropriate nonlinear mapping two classes can be separated with the help of support vectors and margins defined by the support vectors. Training of SVM is extremely slow, but is very accurate due to their ability to model nonlinear decision boundaries.

2. Related works

Education being one of the most important factors that affects the growth of the society is getting developed through different research related programmes. Mining in educational environment is called Educational Data Mining and it's a promising area of research. Data mining in higher education is a recent research field and is gaining popularity because of its potentials to educational institutes. A number of studies have been carried out on the application of data mining techniques in educational purposes.

U. K. Pandey and S. Pal [7] conducted study on the student.Performance based by selecting 600 students from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will perform or not.

S.T. Hijazi and R.S.M.M. Naqvi [8] conducted a study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that was stated as "Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance" was framed. By means of simple linear regression analysis, it was found that the factors like mothers' education and students' family income were highly correlated with the student academic performance

Z. N. Khan [9] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Al-Radaideh et al [10] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the NaïveBayes wereused. The outcome of their results indicated that Decision Tree model had better prediction than other models.

Paulo Cortez and Alice Silva [11] performed a prediction of secondary students' grades of two core classes (Mathematics and Portuguese) by using past school grades (first and second periods), demographic, social and other school related data. Three different data mining goals (i.e. binary/5- level classification and regression) and four data mining methods, i.e. Decision Trees, Random Forests, Neural Networks and Support Vector Machines were tested.The simulation result showed that a good predictive accuracy can be achieved, provided that the first and/or second school period grades are made available.

3. Dataset description

This data approach student achievement in secondary education of two Portuguese schools. We have taken this dataset from the University of California at Irvine (UCI) Machine Learning Repository [12]. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modelled under binary/five-level classification and regression tasks.The target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades.

Here our task to evaluate the final grade of the students in any institute using Multilayer Perceptron (MLP), which is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. Here we train the datasets using a classifier to obtain the most accurate result. Here number of instances are 649 and number of attributes are 33. Our challenge to converting all attribute values to integer type for classification. And no missing values present in the datasets. The information is given in Table 1 below.

Table 1. Dataset description

Serial number	Attributes	Value type	Attribute description
1	School	Binary	Student’s school
2	Sex	Binary	Student’s sex
3	Age	Numeric	Student’s age
4	Address	Binary	Student’s home address type
5	Famsize	Binary	Family size
6	Pstatus	Binary	Parent’s cohabitation status
7	Medu	Numeric	Mother’s education
8	Fedu	Numeric	Father’s education
9	Mjob	Nominal	Mother’s job
10	Fjob	Nominal	Father’s job
11	Reason	Nominal	Reason to choose this school
12	Guardian	Numeric	Student’s guardian
13	Traveltime	Numeric	Time taken by students to go to school
14	Studytime	Numeric	Weekly study time
15	Failures	Binary	Number of past class failures
16	Schoolsup	Binary	Extra educational support
17	Famsup	Binary	Family educational support
18	Paid	Binary	Extra paid classes within the course subject
19	Activities	Binary	Extra-curricular activities
20	Nursery	Binary	Attended nursery school
21	Higher	Binary	Wants to take higher education
22	Internet	Binary	Internet access at home
23	Romantic	Binary	With a romantic relationship
24	Famrel	Numeric	Quality of family relationships
25	Freetime	Numeric	Free time after school
26	Goout	Numeric	Going out with friends
27	Dalc	Numeric	Workday alcohol consumption
28	Walc	Numeric	Weekend alcohol consumption
29	Health	Numeric	Current health status
30	Absences	Numeric	Number of school absences
31	G1	Numeric	First period grade
32	G2	Numeric	Second period grade
33	G3	Numeric	Final grade

4. Proposed method

The detailed procedure is divided into two major steps- data preprocessing followed by data classification. The preprocessing procedure may involve different techniques such as data cleaning and data transformation. In the event of classification, firstly the mathematical model of the classifiers is initialized with default control parameters. After initialization, theyare trained using the training tuples. After the training phase, they are tested with unknown tuples as test input to obtain predicted class label. This label is compared with the actual class label to estimate the accuracy of the classifier being used. The following figure depicts the proposed methodology of the system usingdifferent classifiers.

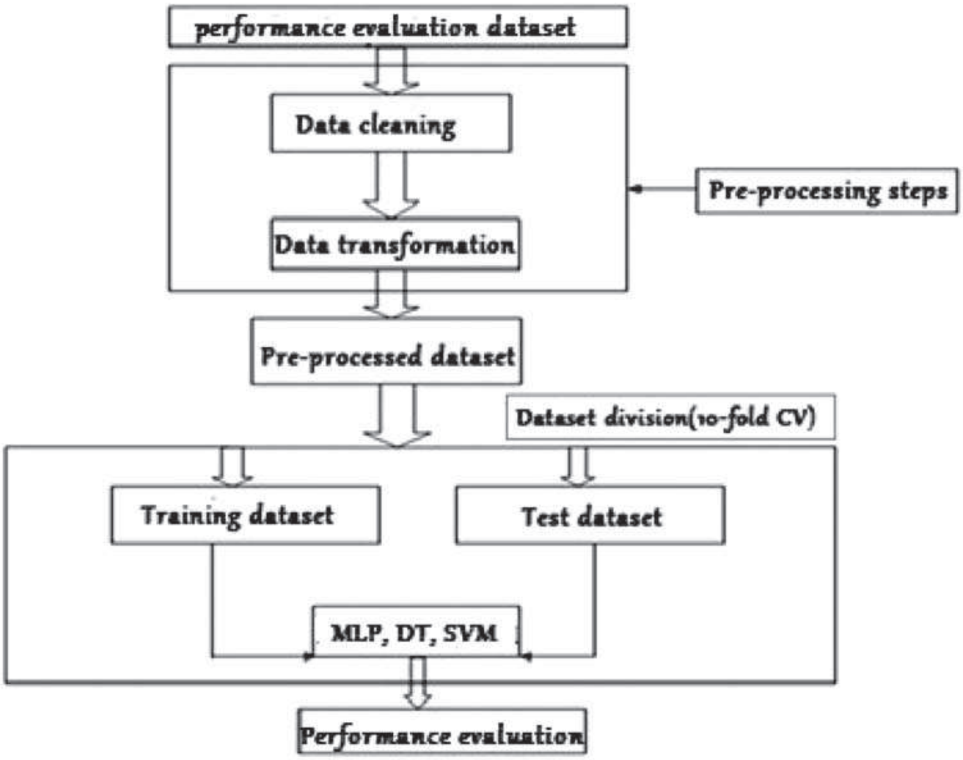

Figure 1: Proposed methodology of the system using different classifiers

The procedure is described in details:

Step 1: Data Preprocessing

Initially, the following data preprocessing techniques are applied to the dataset before the classification task —

Data cleaning: Data cleaning is one of the most important steps to be considered while considering classification of the dataset. Data cleaning makes an attempt to fill in missing values, smoothening of the noise present in the dataset and also correcting the inconsistency present in the dataset. A missing value is normally substituted by the arithmetic mean for that attribute based on statistics. Since there are no missing or inconsistent values in our dataset, this step is not required in our project.

Data transformation: Using this way the dataset is normalized as because the ANN based technique requires distance measurements in the training phase. It converts attribute values to a small-scale range like -1.0 to +1.0.

Step 2: Data classification

Afterwards, the student evaluation dataset is distributed into two disjoint sub-sets, namely the training set and the test set. Basically we employ 10-fold cross-validation for distributing the training and test datasets separately.In the present work, classification technique namely Multi-layer Perceptron (MLP) is trained and tested on the benchmark student evaluation databases.

5. Results and Discussion

Multilayer Perceptron, Decision Tree and Support Vector Machine classifiers are applied to the UCI machine learning repository data set for investigation and performance analysis. We have divided data set into two parts one for training purpose and the other one for testing purpose. The results described here are exclusively based on the simulation experiment that we have taken.Several comparisons are performed; a comparison of classification accuracy, root-mean-square error (RMSE) [13], kappa statistic [14] values and a comparison of True Positive Rate (TP-Rate) or Recall, False Positive Rate (FP-Rate), Precision and F-Measure values derived from the confusion matrix [15] of each classifier.

After the training phase is over, each of the three classifiers(MLP,DT, and SVM) is applied to a test set for classification. Firstly, the performance comparisons of these classifiers are done based on the different performance measures such as accuracy (or classification accuracy), RMSE, and kappa statistic measure as shown below in Table 2.

Table 2. Performance comparisons of three classifiers

Classifier	Accuracy(%)	RMSE	Kappa statistic
MLP	91.2%	0.0773	0.9019
DT	85.5%	0.0985	0.8384
SVM	84.2%	0.1047	0.8135

From Table 2 we could see that, the MLP classifier has an accuracy of 91.2%. Decision Tree classifier has classification accuracy of 85.5%; while SVM classifier has an accuracy of 84.2%. Surely, accuracy wise MLP has performed better than SVM and Decision Tree. Thenwe have analyzed the performance of each classifier based on the information on RMSE and the kappa statistic values collected from Table 2.

Based on the result, MLP comes out first with an RMSE value of 0.0773 and a kappa statistic value of 0.9019; followed by Decision Tree having an RMSE value of 0.0985 and a kappa statistic value of 0.8384 and SVM stands last with the highest RMSE value (0.1047) and the lowest kappa statistic value (0.8135). Therefore, with regard to the performance measures such as classification accuracy, RMSE and kappa statistic, the proposed MLP classifier has performed the best.

Next, the performances of these models are compared based on the TP-Rate (or Recall), FP-Rate, Precision, and F-Measure values derived from the confusion matrix of individuals with respect to the test data set. The detailed accuracy for these classifiers is shown below in Table 3. The weighted average values are also shown in the following table. The results reported here are entirely based on simulation experiment. For evaluating the performance of a classifier, we would expect higher values for TP-Rate, Precision, Recall, F-Measure; and lower valuefor FP-Rate. We have compared the performance of each classifier basedon the information on a weighted average of different performancemeasures from Table 3.

Table 3. Detailed Accuracy for different classifiers

Classifier	TP-Rate(Recall)	FP-Rate	Precision	F-Measure
MLP	91.2%	1.0%	91.1%	91.0%
DT	85.5%	1.6%	85.3%	85.3%
SVM	84.2%	2.4%	84.2%	84.2%

From Table 3 we could discover that the weighted average values of TP-Rate (or Recall), FP-Rate, Precision, and F-Measure for proposed MLP classifier are 91.2%, 1.0%, 91.1%, and 91.0%, respectively; whereas for Decision Tree classifier the values are 85.5%, 1.6%, 85.3%, and 85.3% respectively. For SVM these values are 84.2%, 2.4%, 84.2%, and 84.2% respectively. Surely, the MLP model has the highest weighted average values for TP-Rate (or Recall), Precision, and F-Measure and the lowest weighted average value for FP-Rate. Regarding F-Measure as the best performance measure derived from a confusion matrix; MLP has the highest value for the F-Measure as 91%, followed by Decision Tree having an F-Measure value of 85.3% and SVM with an F-Measure value of 84.2%. The present work uses the 10-fold cross-validation (CV) technique to distribute the training and test datasets foravoiding biases. In other words, the generated training and test datasets are entirely disjoint. So, the classification methods used here can not suffer from the problem of overfitting. We have used some standardmeasures to evaluate the performance of the proposed classification model. For example, well-known evaluation measures like kappa statistic, TP--Rate (or Recall), FP-Rate, Precision, F-Measure are selected here for classification performance analysis. That is why our work is better than this work [11].

6. Conclusion

Prediction of students' performance is most likely one of the challenging subject areas to deal with since so many techniques, tests and data sets are continually being updated, added to or tried out. As a conclusion, we have taken on our objective which is to evaluate and investigate MLP classification algorithms. This study has shown the potential of the artificial neural network for predicting the performance of the students. The different classifiers that we have used are MLP, DT and SVM and we compare our methodology using these traditional classification techniques. The result shows that MLP performs significantly (accuracy is more than 5%) better than the other classification methods used here.

References:

[1] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann, 2000.

[2] A. K. Pujari, Data Mining Techniques Universities Press (India) Private Limited. 1st Edition, 2001.

[3] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, 2nd Edition, 1998.

[4] J. R. Quinlan. Simplifying decision trees International Journal of Man-Machine Studies vol. 27, no. 3, pp. 221234, 1987.

[5] L. Breiman, J. H. Freidman, R.A. Olshen and C. J. Stone. Classification and Regression Trees Belmont, Wadsworth, 1984.

[6] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995.

[7] U. K. Pandey, and S. Pal, “Data Mining: A prediction of performer or underperformer using classification”, (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.

[8] S. T. Hijazi, and R. S. M. M. Naqvi, “Factors affecting student’s performance: A Case of Private Colleges”, Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

[9] Z. N. Khan, “Scholastic achievement of higher secondary students in science stream”, Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.

[10] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, “Mining student data using decision trees”, International Arab Conference on Information Technology-(ACIT'2006), Yarmouk University, Jordan, 2006.

[11] Paulo Cortez and Alice Silva, Using data mining to predictSecondary School student performance, p. 5-12, April 2008.

[12] UCI Dataset URL: https://archive.ics.uci.edu/ml/datasets/student+performance

[13] J. Scott Armstrong and Fred Collopy, Error measures for generalizing about forecasting methods: Empirical Comparisons International Journal of Forecasting, vol. 8: 6980, 1992.

[14] Jean Carletta, Assessing agreement on classification tasks: the kappa statistic Computational Linguistics, MIT Press Cambridge MA, USA, vol. 22, no.2, pp. 249254, 1996.

[15] Stephen V. Stehman, Selecting and interpreting measures of thematic classification accuracy Remote Sensing of Environment, vol. 62, no.1, pp.7789, 1997.