International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

Classification of Heuristic Information by Using Machine Learning Algorithms

Murat KÖKLÜ*[1], Kadir SABANCI[2],Muhammed Fahri ÜNLERŞEN[3]

Accepted 3rd September 2016

Abstract:The User Knowledge Modelling dataset in the UCI machine learning repository was used in this study. The students were classified into 4 class (very low, low, middle, and high) due to the 5 performance data in the dataset. 258 data of 403 data in the dataset were used for training and 145 of them were used for tests. The Weka (Waikato Environment for Knowledge Analysis) software was used for classification. In classification Multilayer Perceptron (MLP), k Nearest Neighbors (kNN), J48, NativeBayes, BayesNet, KStar, RBFNetwork and RBFClassifier machine learning algorithms were used and success rates and error rates were calculated. In this study 8 different data mining algorithm were used and the best classification success rate was obtained by MLP. With Multilayer perceptron neural network model the classification success rates was calculated when there are different number of neurons in the hidden layer of MLP. The best classification success rate was achieved as 97.2414% when there was 8 neurons in the hidden layer. MAE and RMSE values were obtained for this classification success rate as 0.0242 and 0.1094 respectively.

Keywords:*Machine learning, Weka, MLP, kNN, J48*

1. Introduction

Data mining is a multidisciplinary field of computer science. It is the method of autonomously exploring large data sets to reveal patterns and changes that surpass basic scrutiny. Data mining uses complicated numerical methods to divide the data into slices and estimate the next anticipation values. Knowledge Discovery in Data (KDD) is another name of Data mining [1].

Educational data mining (EDM) is described as the scientific audit field focused on the advancement of methods to expose discoveries within the unique types of information that come from academically contexts, and employing these methods to understand students much more better and the contexts that are learnt [2]. Many studies in the literature have been proposed to explore the relationship between successes of the students and his/her culture, habits, life style, family structure etc.

Superby et al. (2006) have search a relation between Academic failures and increase the number of debates among first-year university students. They aims to classify students into three groups: The LOW RISK students; the probability of succeeding of these students is high. The MEDIUM RISK students; these students are the succeed oneswho thanks to the precautions of the university. The HIGH RISK students, the probability of failing of these kind of students is high. It is proved that the most important attributes related with academic performance have been ensured in all the answers that get from 533 students whose first-year in university while 2003 - 2004 academic year's November month. For estimation of academic success the neural networks, discriminant analysis, decision trees and random forests methods were employed [3].Vera et al. (2012) have proposed a genetic algorithm. A novel data mining method to clarify these kind of problems has been proposed. The dataset used in the study was obtained from high school students who are educated in Zacatecas, Mexico. To attain more understandable and efficient rules for classification, genetic algorithm model and different white box methods were compared. It is shown that operations like attribute selection, effective classification and data balancing are so useful to improve truthfulness [4]. Sen et al. (2012) proposed methods to estimate Secondary Education Transition System (SETS) test results in Turkey. By using the big and rich featured data set obtained from SETS the sensitivity analysis have been employed on those prediction modes. The Logistic Regression, Artificial Neural Networks and Support Vector Machines methods have been employed for predicting. The results were discussed [5].

In this study, the User Knowledge Modelling dataset obtained from UCI Machine Learning Repository have been used for classification. In this dataset, there are 5 attributes that used for determining the students' educational status in 4 class as very low, low, middle, high. For classification 8 machine learning algorithm have been used. The obtained success rates and error values by those algorithms, have been compared.

2. Material and Methods

2.1. Software-WEKA

Weka (Waikato Environment for Knowledge Analysis) written in Java, developed at the University of Waikato, New Zealand [6].

[1] *Selcuk University, Faculty of Technology, Computer Engineering, Konya, Turkey*
[2] *KaramanogluMehmetbey University, Faculty of Engineering, Department of Electrical and Electronics Engineering, Karaman, Turkey.*
[3] *Necmettin Erbakan University,Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, Konya, Turkey.*
** Corresponding Author:Email:mkoklu@selcuk.edu.tr*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

Weka provides a lot of standard data mining methods. More of them, pre-processing of data, clustering, classification, regression, visualization, and selection or attributes.The thought that all techniques of Weka's software are based on is that the data is available as a single flat file or relation, where a fixed number of attributes identify each data point (normally, numeric or nominal attributes, but some other attribute types are also supported) [7].

2.2. k-Nearest Neighbour Algorithm

A supervised learning algorithm, k-NN solves classification problems. Classification is the examination of the attributes of an image and the designation of this image to a predefined class. The critical point is the determination of the features of each category previously [8]. Conforming to the used classifcation algorithm k-NN based on the attributes drawn from the classification stage, the distance of the new individual that is wanted to be classified to all previous individuals is considered and the nearest k class is used. As an outcome of this procedure, the belonging of the test data is determined due to the k-nearest neighbour category which contains more exactly determined classes. In k-NN, the determination of the algorithm used for distance calculation and neighbour number are the critical optimization points. In the study, the optimum k number is appointed with experiments. In the calculation of distance, the Euclidean Distance is performed.

Euclidean calculation method [9]:

$$d(x_i, x_j) = \left(\sum_{s=1}^{p} (x_{is} + x_{js})^2 \right)^2$$

xi and xj are two points that is wanted to be learnt the distance between them.

2.3. Multilayer Perceptron

It is a feed forward type artificial neural network model which corresponds input sets onto proper sets of output. A multilayer perceptron (MLP) is composed of multiple layers where each layer is connected to the other one. Each node is a processing element or a neuron that has a nonlinear activation function except the input nodes. It uses a supervised learning technique named back propagation and it is used for training the network. The alteration of the standard linear perceptron, MLP is capable of distinguishing data which are not linearly separable [6].

2.4. RBF Network

RBF Network is a kind of artificial neural network. The difference is in its activation functions. The activation function used in RBF Networks is radial basis functions. Results of inputs and neuron parameters by using radial basis functions are combinated to obtain the output of the network. In the applications such as estimation, classification, time series and system controls the radial basis function networks could be employed [6].

2.5. BayesNet

What represents a number of random variables and their conditional dependencies by way of a directed acyclic graph (DAG) is stochastic graphical model (a type of statistical model). For example, a Bayesian network could stand for the stochastic relationships between diseases and symptoms. If symptoms are given, in order to estimate the probabilities of the existence of various diseases, the network can be used [6].

2.6. Naïve Bayes

Naive Bayes classifiers needing some parameters linear in the number of factors (features/predictors) in a learning problem are extremely scalable. The evaluation of a closed-form expression taking linear time, instead of expensive iterative approximation as used for a number of other kinds of classifiers provide maximum-likelihood training [6].

2.7. J48

J48 algorithm of Weka software attributed to J.R. Quilan C4.5 algorithm is a famous machine learning algorithm. All data get to be studied will be of the categorical type and so continuous data will not be studied at this stage. In order to contain this talent, the algorithm will however leave room for adaption to contain this talent [6, 7].

2.8. KStar

K-star or K* is an instance-based classifier. The class of a test example is attributed to the training samples similar to it, as specified by some similarity function. In terms of using an entropy-based distance function, it differs from other instance-based learners. [10].

3. Results and Discussion

403 sample data at User Knowledge Modelling dataset were processed by using Weka program. The classification accomplishment, MAE and RMSE valuesfor different number of neuronsin the hidden layers were obtained. The diagram demonstrating the changes in MAE and RMSE values due to the number of neurons in the hidden layer, is demonstrated in Figure 1.

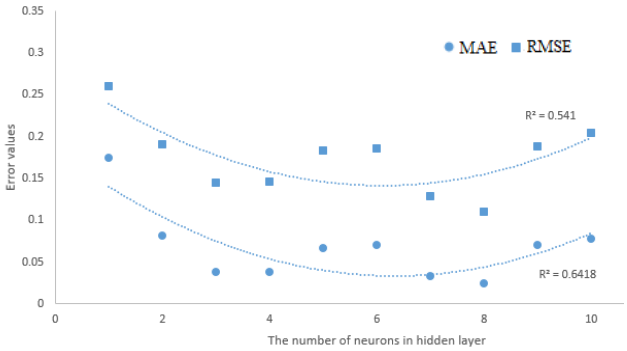

Figure 1. Change of error valueswiththe number of neurons in hidden layer

The classification accomplishmentwas obtained by using kNN algorithm for same dataset. By using k-NN method, classification accomplishments were obtained for different k neighborhood values. Additionally, the mean absolute error (MAE) values and the root mean square error (RMSE) values were calculated. The diagram presenting the change of MAE and RMSE values with the number of neighborhood in k-NN algorithm is shown in Figure 2.

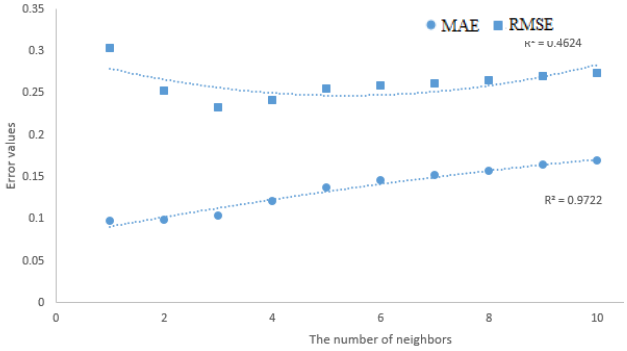

Figure 2. Change of error valueswith the number of neighborhood

Then the same data was processed using J48, NaiveBayes, RBFNetwork, RBFClassifier, BayesNet, Kstar machine learning algorithms and classification accomplishments, MAE and RMSE values of different tree types in the forest were obtained. The success and error rates obtained using 8 different classification algorithms (Multilayer Perceptron, kNN, J48, Naïve Bayes, Bayes Net, KStar) can be seen in Table I. The diagram demonstrating the error values obtained based on different machine learning algorithms can be seen in Figure 3.

Table 1. The ClassificationAccomplishments Obtained By Using Various Machine Learning Algorithms

The number of neurons in the hidden layer	*Classification Success (%)*	*MAE*	*RMSE*
MLP	97.2414	0.0242	0.1094
kNN	87.5862	0.1037	0.2324
J48	91.0345	0.0481	0.2092
NaiveBayes	84.8276	0.1173	0.2582
RBFNetwork	93.7931	0.0411	0.1567
RBFClassifier	71.0345	0.2143	0.2988
BayesNet	86.8966	0.1312	0.2485
Kstar	81.3793	0.1122	0.2695

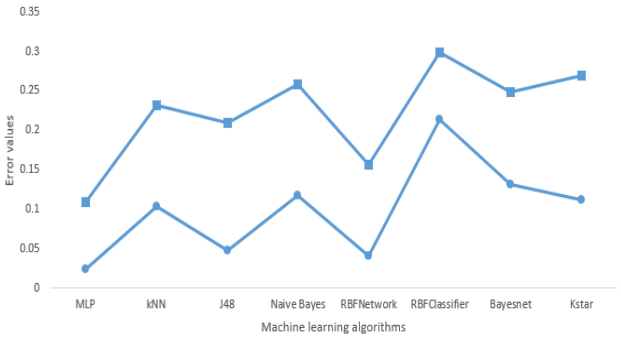

Figure 3. Variation of error rate based on the machine learning algorithms

4. Conclusion

In this study, students have been classified about their success in the school as very low, low, middle and high by using their daily habits. By this intention, popular data mining algorithms like kNN, MLP, J48, NaiveBayes, RBFNetwork, RBFClassifier, BayesNet and Kstar, have been used and compared with each other. At attained classification accomplishments, success rates has better while using with k-NN algorithm. At classification accomplishments, obtained by using k-NN algorithm, the best classification success rate was occurred for 3 neighbourhood as 87.5862%. At this neighbourhood, MAE and RMSE values are 0.1037 and 0.2324 respectively. In this study, classification accomplishments obtained with Multilayer perceptron algorithm are very low when compared with k-NN classifier. The highest classification accomplishment of Multilayer perceptron algorithm is 97.24% when there are 8 neurons in its hidden layer. For this situation the MAE and RMSE values are 0.0242 and 0.1094 respectively.

The success rates obtained using J48, NaiveBayes, RBFNetwork, RBFClassifier, BayesNet and Kstar classification algorithms were found as 91.0345%, 84.8276%, 93.7931%, 71.0345%, 86.8966% and 81.37935.respectively.

References

[1] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[2] B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.

[3] Superby J. F., Vandamme J. P., Meskens N. Determination of factors influencing the achievement of the first-year university students using data mining methods. In International conference on intelligent tutoring systems, Educational Data Mining Workshop, Taiwan, 2006:1 – 8.

[4] Márquez-Vera, C., Cano, A., Romero, C., Ventura, S. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data, Applied Intelligence, April 2013, Volume 38, Issue 3, pp 315-330.

[5] Sen B., Ucar E., Delen D. Predicting and analyzing secondary education placement-test scores: A data mining approach, Expert Systems with Applications, Vol. 39, No. 10, pp. 9468-9476, 2012.

[6] WEKA, http://www.cs.waikato.ac.nz/~ml/weka/ Last access: 10.04.2015.

[7] Rohit Arora and Suman, Comparative Analysis of Classification Algorithms on Different Datasets using WEKA, International Journal of Computer Applications (0975 – 8887) Volume 54– No.13, September 2012.

[8] J. Wang, P. Neskovic, and L. N. Cooper, "Improving nearest neighbor rule with a simple adaptive distance measure", Pattern Recognition Letters, 28(2):207-213, 2007

[9] Y. Zhou, Y. Li, and S. Xia, "An improved KNN text classification algorithm based on clustering", Journal of computers, 4(3):230-237, 2009.

[10] John G. Cleary, Leonard E. Trigg: "K*: An Instance based Learner Using an Entropic Distance Measure", 12th International Conference on Machine Learning, 108-114, 1995.