Performance Analysis of Data Mining Techniques for Placement Chance Prediction

V.Ramesh, P.Parkavi, P.Yasodha

Abstract- Predicting the performance of a student is a great concern to the higher education managements. The scope of this paper is to investigate the accuracy of data mining techniques in such an environment. The first step of the study is to gather student's data. We collected records of 300 Under Graduate students of computer science course, from a private Educational Institution. The second step is to clean the data and choose the relevant attributes. In the third step, NaiveBayesSimple, MultiLayerPerception, SMO, J48, REPTree algorithms were constructed and their performances were evaluated. The study revealed that the MultiLayerPerception is more accurate than the other algorithms. This work will help the institute to accurately predict the performance of the students.

Keywords: Data Mining, Classification, Decision Tree Algorithm, placement Prediction

1. INTRODUCTION

IN real world, predicting the performance of the students is a challenging task. The primary goals of Data Mining in practice tend to be Prediction and Description [1]. Predicting performance involves variables like Maths, Programming language, Lab Marks etc. in the student database to predict the unknown or future values of interest. Educational Data Mining uses many techniques such as Decision Trees, MultiLayerPerception, NaïveBayes and many others. Using these methods many kinds of knowledge can be discovered.

The main objective of this paper is to use data mining methodologies to study students' performance in the courses. Data mining provides many tasks that could be used to study the student performance. In this research, the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree and Naïve Bayes, MultiLayerPerception method is used here. Information's like English, Maths, Programming language, Lab Marks , Placement Details were collected from the existing database. Marks were collected from the Department of Computer science, to predict their placement performance at the end of the Final semester. This paper investigates the accuracy of NaiveBayesSimple, MultiLayerPerception, SMO, J48, REPTree, techniques for predicting student performance.

2. RELATED WORK

In (Galit, 2007) gave a case study that use students data to analyze their learning behavior to predict the results and to warn students at risk before their final exams.

Cortez and Silva [6] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school students from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms such as Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Support Vector Machine (SVM) were applied on a data set of 788 students, who appeared in 2006 examination. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (pass/fail) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset.

(Erdogan and Timor 2005) used educational data mining to identify and enhance educational process which can improve their decision making process. Finally (Henrik ,2001) concluded that clustering was effective in finding hidden relationships and associations between different categories of students.

Khan [2] 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.

Modeling of student performance at various levels is discussed in [4], [5], and [6]. Ma, Liu, Wong, Yu, and Lee [4] applied a data mining technique based on association rules to find weak tertiary school students (n= 264) of Singapore for remedial classes. Three scoring measures namely Scoring Based on Associations (SBA-score), C4.5-score and NB-score for evaluating the prediction in connection with the selection of the students for remedial classes were used with the input variables like sex, region and school performance over the past years. It was found that the predictive accuracy of SBA-score methodology was 20% higher than that of C4.5 score, NB-score methods and traditional method.

Kotsiantis, et al. [5] applied five classification algorithms namely Decision Trees, Perceptron-based Learning, Bayesian Nets, Instance-Based Learning and Rule-learning to predict the performance of computer science students from distance learning stream of Hellenic Open University, Greece. A total of 365 student records comprising several demographic variables like sex, age and marital status were used. In addition, the performance attribute namely mark in a given assignment was used as input to a binary (pass/fail) classifier. Filter based variable selection technique was used to select highly influencing variables and all the above five classification models were constructed. It was noticed that the Naïve-Bayes yielded high predictive accuracy (74%) for two-class (pass/fail) dataset.

3. METHODOLOGY

The methodology was adopted to generate a database for the current study. For this study we collected secondary data. The data were collected from the computer science department and the placement cell from a college.

3.1. DATA SOURCE

A sample of 300 student's record was taken from a computer science department of a college. The attributes used are presented in table with the values of every attribute. In which the English, Maths, Programming language, Practical marks were collected from the department and their Placement details were collected from the placement cell.

The seven semesters Maths, Programming, Practical marks are converted as a single Maths Programming and Practical marks

Before processing data we are going to clean the data to remove noise and inconsistent. To remove missing values in the dataset we use the cleaning technique.

VARIABLE NAME	DESCRIPTION	VALUES
ENGLISH	student english mark	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
MATHS-1	student maths mark in first semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
MATHS-3	student maths mark in third semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
MATHS-4A	student maths mark in fourth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
MATHS-4B	student maths mark in fourth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
MATHS-5	student maths mark in fifth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
MATHS-6	student maths mark in sixth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
MATHS-7	student maths mark in seventh semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
PRG-1	student programming language mark in first semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
PRG-3	student programming language mark in third semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
PRG-4	student programming language mark in fourth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
PRG-5A	student programming language mark in fifth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
PRG-5B	student programming language mark in fifth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
PRG-6	student programming language mark in sixth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
LAB-1	student lab mark in first semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}

LAB-3	student lab mark in third semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
LAB-5A	student lab mark in fifth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
LAB-5B	student lab mark in fifth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
LAB-6	student lab mark in sixth semester	{S-10,A-9,B-8,C-7,D-6,E-5,F4}
PLACEMENT	student placement detail	{YES,NO}

3.2. TOOLS AND TECHNIQUES

Selecting a Data Mining Tool

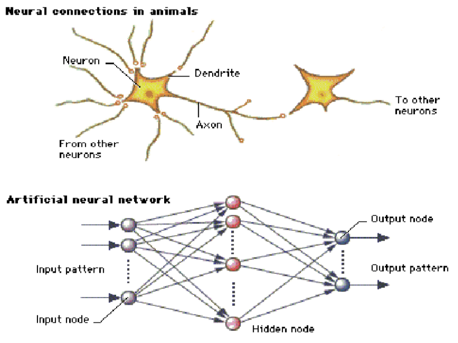
Hence, this paper proposes to use an open source data mining tool-Weka. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato. WEKA is free software available under the GNU General Public License. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. However, as is mentioned above, Weka is an open source data mining tool which can be extended by the users, that helps users a lot, when tools Weka provides that can not meet the users requirement, they can develop new tool kits and add them to Weka. Therefore, Weka is a very good data mining tool which could be used in the field of education. In that we are going to use classification technique. The classification techniques of data mining help to classify the data on the basis of certain rules. This helps to frame policies for the future.

NEURAL NETWORK

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.

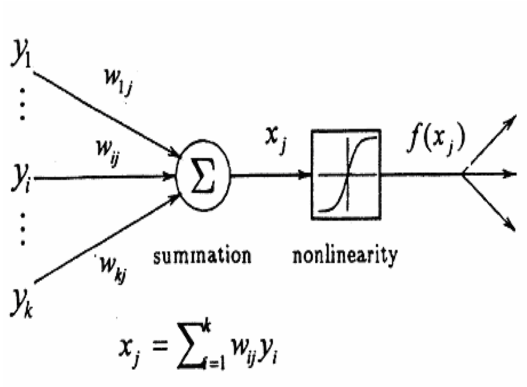
The Biological Model

Artificial neural networks emerged after the introduction of simplified neurons by McCulloch and Pitts in 1943 (McCulloch & Pitts, 1943). These neurons were presented as models of biological neurons and as conceptual components for circuits that could perform computational tasks. The basic model of the neuron is founded upon the functionality of a biological neuron. "Neurons are the basic signalling units of the nervous system" and "each neuron is a discrete cell whose several processes arise from its cell body".



Mathematical Model

When creating a functional model of the biological neuron, there are three basic components of importance. First, the synapses of the neuron are modelled as weights. The strength of the connection between an input and a neuron is noted by the value of the weight. Negative weight values reflect inhibitory connections, while positive values designate excitatory connections [Haykin]. The next two components model the actual activity within the neuron cell. An adder sums up all the inputs modified by their respective weights. This activity is referred to as linear combination. Finally, an activation function controls the amplitude of the output of the neuron. An acceptable range of output is usually between 0 and 1, or -1 and 1.



Neural networks architectures

An ANN is defined as a data processing system consisting of a large number of simple highly inter connected processing elements (artificial neurons) in an architecture inspired by the structure of the cerebral cortex of the brain. There are several types of architecture of NNs.

Feed-forward neural networks

In a feed forward network, information flows in one direction along connecting pathways from the input layer via the hidden layers to the final output layer. There is no feedback (loops) i.e., the output of any layer does not affect that same or preceding layer.

Recurrent neural networks

These networks differ from feed forward network architectures in the sense that there is at least one feedback loop. Thus, in these networks, for example, there could exist one layer with feedback connections as shown in figure below. There could also be neurons with self- feedback links, i.e. the output of a neuron is fed back into itself as input.

Learning/Training methods

Learning methods in neural networks can be broadly classified into three basic types:
supervised, unsupervised and reinforced

Supervised learning

In this, every input pattern that is used to train the network is associated with an output pattern, which is the target or the desired pattern. A teacher is assumed to be present during the learning process, when a comparison is made between the network's computed output and the correct expected output, to determine the error. The error can then be used to change network parameters, which result in an improvement in performance.

Unsupervised learning

In this learning method, the target output is not presented to the network. It is as if there is no teacher to present the desired patterns and hence, the system learns of its own by discovering and adapting to structural features in the input patterns.

Reinforced learning

In this method, a teacher though available, does not present the expected answer but only indicates if the computed output is correct or incorrect. The information provided helps the network in its learning process. A reward is given for a correct answer computed and a penalty for a wrong answer. But, reinforced learning is not one of the popular forms of learning.

The Back Propagation Algorithm

Back propagation, or propagation of error, is a common method of teaching artificial neural networks how to perform a given task. The back propagation algorithm is used in layered feed forward ANNs. This means that the artificial neurons are organized in layers, and send their signals "forward", and then the errors are propagated backwards. The back propagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The idea of the back propagation algorithm is to reduce this error, until the ANN learns the training data.

Summary of the technique

1. Present a training sample to the neural network.
2. Compare the network's output to the desired output from that sample. Calculate the error in each output neuron.
3. For each neuron, calculate what the output should have been, and a scaling factor, how much lower or higher the output must be adjusted to match the desired output. This is the local error.
4. Adjust the weights of each neuron to lower the local error.
5. Assign "blame" for the local error to neurons at the previous level, giving greater responsibility to neurons connected by stronger weights.
6. Repeat the steps above on the neurons at the previous level, using each one's "blame" as its error.

NAIVEBAYES CLASSIFIER

A NaiveBayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

MULTILAYER PERCEPTION

A multilayer perceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate output. 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 called backpropagation for training the network. MLP is a modification of the standard linear perceptron

Confusion Matrix

The confusion matrix is more commonly named contingency table. In our case we have two classes, and therefore a 2x2 confusion matrix. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified. For example the confusion matrix of MLP is shown below.

```
=== Confusion Matrix ===
 a     b  <-- classified as
75     6 |  a = YES
 9    29 |  b = NO
```

J48 & SMO

J48 is an implementation of C4.5 release 8 [7] that produces decision trees. This is a standard algorithm that is widely used for practical machine learning. Part is a more recent scheme for producing sets of rules called "decision lists"; it works by forming partial decision trees and immediately converting them into the corresponding rule. SMO implements the "sequential minimal optimization" algorithm for support vector machines, which are an important new paradigm in machine learning [9].

REPTREE

Fast decision tree learner. Builds a decision/regression tree using information gain/variance reduction and prunes it using reduced-error pruning (with backfitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5). The table below describes the options available for REPTree.

Option	Description
Debug	If set to true, classifier may output additional info to the console.
maxDepth	The maximum tree depth (-1 for no restriction).
minNum	The minimum total weight of the instances in a leaf.
minVarianceProp	The minimum proportion of the variance on all the data that needs to be present at a node in order for splitting to be performed in regression trees.
noPruning	Whether pruning is performed.
numFolds	Determines the amount of data used for pruning. One fold is used for pruning, the rest for growing the rules.
Seed	The seed used for randomizing the data.

REPTree does reduced-error pruning.

4. RESULTS

This research is a starting attempt to use data mining functions to analyze and evaluate student academic data and to enhance the quality of the educational system.

We tested data set with five different classification algorithms: NaiveBayesSimple, Multilayer Perception, SMO, J48, REPTree . In our study we are going to compare the correctly classified instances as well as mean absolute error with different algorithms they are, Naive Bayes Simple, Multilayer Perception, SMO, J48, and REPTree. The following table shows this comparison.

Algorithm	Correctly classified Instances	Mean Absolute Error
NaïveBayesSimple	83.1933%	0.1852
MultilayerPerception	87.395%	0.2002
SMO	84.0336 %	0.1597
J48	84.8739%	0.2553
REPTree	84.8739%	0.231

This work will help the institute to accurately predict the performance of the students and to find out the weak student. This will help to improve the performance of such student in the early stage.

5. CONCLUSION

This paper illustrates how well different classification techniques are used as predictive tools in the data mining domain and after comparing their performances. From the results it is proven that MultiLayerPerception algorithm is most appropriate for predicting student performance. MLP gives 87% prediction which is relatively higher than other algorithms. This study is an attempt to use classification algorithms for predicting the student performance and comparing the performance of NaiveBayesSimple, MultiLayerPerception, SMO, J48, and REPTree.

6. REFERENCES

[1] David Hand, Heikki Mannila, Padhraic Smyth "Principles of Data Mining"

[2] Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream", Journal of Social Sciences, Vol. 1, No. 2, 2005, pp. 84-87.

[3] Ye zhiwei, Hu zhengbing "Research on application data mining to teaching of basic computer courses in universities"

[4] Y. Ma, B. Liu, C.K. Wong, P.S. Yu, and S.M. Lee, "Targeting the Right Students Using Data Mining", Proceedings of KDD, International Conference on Knowledge discovery and Data Mining, Boston, USA, 2000, pp. 457-464.

[5] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques", Applied Artificial Intelligence, Vol. 18, No. 5, 2004, pp. 411-426.

[6] P. Cortez, and A. Silva, "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp.5-12.

[7] Quinlan, J.R. (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA.

[8] Burges, C.J.C. (1998) "A tutorial on support vector machines for pattern recognition." Data Mining and Knowledge Discovery, Vol. 2(1), 121-167.

[9] Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo(1993)

[10] Weka – Data Mining Machine Learning Software, http://www.cs.waikato.ac.nz/ml/

AUTHOR PROFILE

V.Ramesh received his M.Phil. in the area of Data Mining from Madurai Kamaraj University, Madurai. He is doing PhD degree in the area of Data Mining in Agriculture. At present he is working as Assistant Professor at Department of Computer Applications in Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University, Kanchipuram, Tamil Nadu. He has published more than 5 research papers in National, International journals and conferences. His research interest lies in the area of Data Mining, Artificial Intelligence, Neural Networks, Database Management Systems.

P.Parkavi received her Master Degree from SCSVMV University and currently doing M.Phil Research at SCSVMV University, Kanchipuram. Her research interest lies in the area of Data Mining.

P.Yasodha Mphil Research Scholar in the Department of Computer Science & Applications in Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University, Enathur, Kanchipuram. She received the degree in Master of Computer Applications from SCSVMV University in 2010. At present she is working as Professor at Department of Computer Science in Pachaiyappa's College for Women, kanchipuram, Tamil Nadu. She has published 3 research papers in National, International Journals and conferences . Her research interest lies in the area of Data Mining, Neural Networks.