International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

Performance Evaluation By Artificial Neural Network Using WEKA

Sumam Sebastian

Assistant Professor, Department of Computer Science, BVM Holy Cross College, Cherpunkal, Kerala, India

---

Abstract - ***Data mining*** *is the process of extracting hidden patterns and useful information from large set of data is now becoming part of current inventions. Data mining now can be applied to different fields like marketing, education; health etc. Data mining in field of education is named as educational data mining. Educational data mining can help institutions to predict the performance of their students so as to improve their academic results. In this paper artificial neural network is used to predict the performance of student. Multilayer Perceptron Neural Network is used for the implementation of prediction strategy. The basic concepts og genetic algorithm is applied to the result to obtain better performance.Experiment is conducted using weka and real time dataset available.*

Key Words: **Data Mining; Educational Data Mining; Artificial Neural Network; Multilayer Perceptron Neural Network(MLP);Association Rule Mining;Genetic Algorithm;**

1. INTRODUCTION

Data mining [1] is the process of analyzing data from different perspectives and summarizing it into important information so as to find hidden patterns from a large data set. Data mining [2] points to the strategy of discovering of implicit, previously unknown and practically useful information from the data in the databases. It uses techniques of machine learning, statistical and visualization to discover and present knowledge in a form which is easily understandable to us. The abundance and fast evolution of the data mining discipline comes from its large variety of research areas of interest. Data mining applications adopts different kind of parameters to examine the data. Educational Data Mining[3] is a newly emerged technique that helps to discover methods that will explore unique types of data from education database and helps to predict students' academic performance.

It is very necessary for an institution to maintain a good academic result, for that student's academic performance has to be maintained in better manner. So a continuous student's performance evaluation strategy has to be invented. Different kinds of data mining techniques can be used for this like association rule mining.K-means clustering, artificial neural network etc. Among the different methods most advanced and accurate method is the evaluation using artificial neural network.MLP and the use of genetic algorithm improves the functionality of artificial neural network

2. OBJECTIVES

The main objectives are, first to determine all the personal and academic factors that affects the performance of student, second to transform these factors to a suitable form for system coding and third is to model a neural network that can predict the performance based on the data of student. The main concept used in this paper is that of artificial neural network.

3. THE ARTIFICIAL NEURAL NETWORKS

An artificial neural network (ANN)[4], often called as a "neural network" (NN), is a computational model based on the biological neural networks, in other words, is a representation and emulation of human neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation . In practical terms neural networks are non-linear statistical data modeling tools [5]. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining [6].

Multilayer Perceptron

The most popular form of neural network architecture is the multilayer perceptron (MLP). A multilayer perceptron:

- Has any number of inputs.
- Has one or more hidden layers with any number of units.
- Uses generally sigmoid activation functions in the hidden layers.
- Have connections between the input layer and the first hidden layer, between the hidden layers, and between the last hidden layer and the output layer.

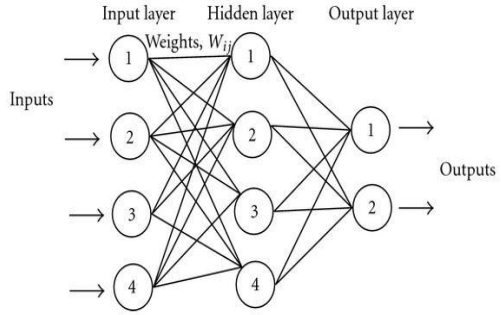

Figure 1: Feed forward neural network

MLP is especially suitable for approximating a classification function which sets the example determined by the vector attribute values into one or more classes. MLP trained with back propagation algorithm is used for data mining.

4. DATA COLLECTION AND PREPROCESSING

In this research paper the data was collected from the two classes 8 and 9 of a school.A dataset of 300 students was used for the evaluation. Neural network is used for predicting the student performance. The attributes selected are mainly of two types, first academic attributes related to the academic details of student and personal attributes related to the personal details of student that affects the study and performance of study of student. The academic attributes selected are 1.Interest of study of the student categorized as low, average and good, 2.Unit test mark the average mark of unit tests conducted divided as low, average and good, 3.Assignment mark which is the average of all assignments divided as average and good, 4.Attendance score which is the average of attendance of the student taken divided into average and good,5.Extracurricular activities performance which is the performance of student in other activities along with studies grouped in to low average and good,6.Residence which is the staying of student categorized into either hostler or day scholar. The personal attributes selected are 1.parent's education and family status where parents education divided as poor average and good and 2.family status is divided as low and average and good. In a given dataset Data Pre-Processing technique is used to identify noise data, missing attribute values, irrelevant and redundant data.

ATTRIBUTES	DESCRIPTION	VALUES
INTEREST OF STUDY	Interest of student in studying	Low Average Good
UNIT TEST MARK	Average mark of student in unit tests	Low Average Good
ASSIGNMENT	Average of marks of assignments	Average Good
ATTENDANCE	Average attendance of student in the class	Average Good
EXTRACARRICULAR ACTIVITIES	Performance of student in extracurricular activities	Low Average Good
RESIDENCE	Residence of student in studying ie in hostel or not	Non Hoster Hostler
PARENT EDUCATION	Education of the parents of student	Poor Average Good
FAMILY STATUS	The total family status of the student	Low Average and good

Table1. Attributes and Its Possible Values

Combining Genetic Algorithms with Neural Networks

In real life, the success of an individual is not only determined by his knowledge and skills, which he gained through experience (the neural network training), it also depends on his genetic heritage (set by the genetic algorithm).One of the genetic factors that affects the student's performance is parents education or family members education and cultural background

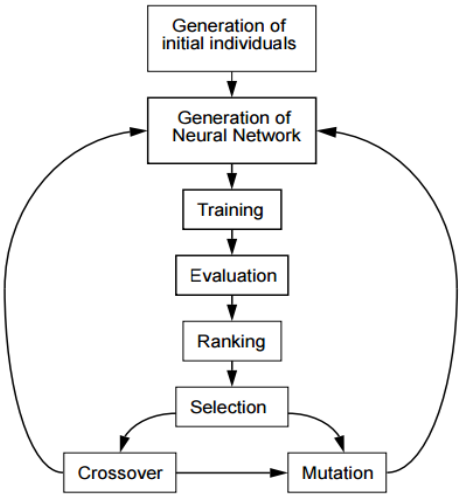

Figure 2: Structure of GANN System

Genetic Algorithms

Genetic algorithms are algorithms for optimization and learning based loosely on several features of biological evolution. They require five components: 1 A way of encoding solutions to the problem on chromosomes. 2. An evaluation function that returns a rating for each chromosome given to it. 3. A way of initializing the population of chromosomes. 4. Operators that may be applied to parents when they reproduce to alter their genetic composition. Included might be mutation, crossover (i.e. recombination of genetic material), and domain-specific operators. 5. Parameter settings for the algorithm, the operators, and so forth.

5. METHODOLOGY

In this research we have used Weka for the entire implementation. Simple training and testing using multilayer perceptron neural network was done first. For this the entire data set was divided into two separate tests. Half for training set and another half for testing set. Training was done by adjusting the different learning and momentum rates. Among the training results the best was taken for analysis. Second MLP training after association rule mining is done. Association rule mining extracts the important rules so that it helps to identify the important attributes. Unnecessary attributes are removed from
the data set and MLP neural network training is done. It gives a better result than simple train & test method. For the most fare evaluation of result K-fold cross validation method of MLP training was used. In this the entire data set is not divided into two different sets as of prior, instead as the input to the system whole data set is given.10 filed cross validation is used here. In this the entire data set is divided into 10 subsets. Among this one set is used as test set and remaining nine sets are used training set.Then genetic algorithm is applied to the result for better performance

5.1WEKA Environment

WEKA [8] stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. WEKA supports many data mining tasks such as data re-processing, classification, clustering, regression and feature selection to name a few.
The supported data formats are ARFF, CSV, C4.5 and binary. Alternatively you could also import from URL or an SQL database. After loading the data, preprocessing filters could be used for adding/removing, attributes, discretization, Sampling, randomizing etc.Weka is a collection of machine learning algorithms for data mining & machine learning tasks. Weka is open source software issued under the GNU General Public License.

5.2 MLP Training with WEKA

Two sets are used for MLP neural network training in WEKA. They are Training set and Test set [9].

 Training set: A set of examples used for learning that is to fit the parameters [i.e., weights] of the classifier.

 Test set: A set of examples used only to assess the performance [generalization] of a fully-specified classifier.

Back Propagation algorithm is used for the network training.

Back Propagation algorithm

Initialize all weights to small random numbers
Until satisfied DO

- For each training example Do
  1. Input the training example to the network and compute the training outputs
  2. For each output unit k

$$\delta_k \leftarrow o_k(1-o_k)(t_k-o_k)$$

  3. For each hidden unit h

$$\delta_h \leftarrow o_h(1-o_h)\sum_{k\in outputs} w_{h,k}\delta_k$$

  4. Update each network weight $w_{i,j}$

$$w_{i,j} \leftarrow w_{i,j} + \Delta w_{i,j}$$

Where

Learning rate

$$\Delta w_{i,j} = \eta \delta_j x_{i,j}$$

Momentum

$$\Delta w_{i,j}(n) = \eta \delta_j x_{i,j} + \alpha \Delta w_{i,j}(n-1)$$

Here we are adjusting the learning rate and momentum to get a better training result.
Main parameters for learning: hiddenLayers, learningRate,momentum,trainingTime(epoh, seed .The parameter setting function is given as
weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

hiddenLayers -- This defines the hidden layers of the neural network. This is a list of positive whole numbers. 1 for each hidden layer. Comma seperated. To have no hidden layers put a single 0 here. This will only be used if autobuild is set. There are also wildcard values 'a' = (attribs + classes) / 2, 'i' = attribs, 'o' = classes , 't' = attribs + classes.
learningRate -- The amount the weights are updated.
momentum -- Momentum applied to the weights during updating.
For fare evaluation, the 'cross-validation' scheme is used

K-fold Cross Validation

- the data set is randomly divided into k subsets.
- One of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set.

5.3 Association Rule Mining

In EDM, association rule learning is a conventional and well researched method for determining interesting relations between attributes in large databases [10]. Association rule Mining is mainly intended to recognize strong rules from databases using different measures of support and confidence.
Support (s) and confidence (c) are two measures of rule interestingness. They truely reflect the usefulness and certainty of the discovered rule.

Apriori Algorithm

Apriori is a seminal algorithm proposed by R. Agarwal and R. Srikant in 1994 for mining frequent item sets for Boolean association rules.
The following lines state the steps in generating frequent item set in Apriori algorithm. [11]
Let Ck be a candidate item set of size k and Lk as a frequent item set of size k. The main steps of iteration are:
• Find frequent set Lk-1
• Join step: Ck is generated by joining Lk -1 with itself (Cartesian product Lk-1 x Lk-1)
• Prune step (apriori property): Any (k − 1) size item set that is not frequent cannot be a subset of a frequent k size item set, hence should be removed
• Frequent set Lk has been achieved [11].
The parameter setting function in weka for association rule mining is
weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
car -- If enabled class association rules are mined instead of (general) association rules.
classIndex -- Index of the class attribute. If set to -1, the last attribute is taken as class attribute.
delta -- Iteratively decrease support by this factor. Reduces support until min support is reached or required number of rules has been generated.

5.4 Genetic algorithm

The basic process for a genetic algorithm is[13]:

1. Initialization - Create an initial population. This population is usually randomly generated and can be any desired size, from only a few individuals to thousands.
2. Evaluation - Each member of the population is then evaluated and we calculate a 'fitness' for that individual. The fitness value is calculated by how well it fits with our desired requirements. These requirements could be simple, 'faster algorithms are better', or more complex, 'stronger materials are better but they shouldn't be too heavy'.
3. Selection - We want to be constantly improving our populations overall fitness. Selection helps us to do this by discarding the bad designs and only keeping the best individuals in the population. There are a few different selection methods but the basic idea is the same, make it more likely that fitter individuals will be selected for our next generation.
4. Crossover - During crossover we create new individuals by combining aspects of our selected individuals. We can think of this as mimicking how sex works in nature. The hope is that by combining certain traits from two or more individuals we will create an even 'fitter' offspring which will inherit the best traits from each of it's parents.
5. Mutation - We need to add a little bit randomness into our populations' genetics otherwise every combination of solutions we can create would be in our initial population. Mutation typically works by making very small changes at random to an individuals genome.
6. And repeat - Now we have our next generation we can start again from step two until we reach a termination condition.

The different classes made for using genetic algorithm concepts are

- Population - Manages all individuals of a population
- Individual - Manages an individuals
- Algorithm - Manages our evolution algorithms such as crossover and mutation
- FitnessCalc - Allows us set a candidate solution and calculate an individual's fitness

6. PERFORMANCE EVALUATION

To evaluate the performance of above methods of neural network training different parameters are available like Accuracy, Precision, Recall, F-Measure, Kappa score etc. Here accuracy, precision and recall are considered.

- Accuracy (percent correct)

Accuracy is how close a measured value is to the actual (true) value. Accuracy retrieves the percentage of correctly classified instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

- Precision

Precision is a value of the accuracy provided by a unique class that was predicted.

$$\text{Precision} = \frac{TP}{TP+FP}$$

 Recall

Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is c also called sensitivity, and points to the true positive rate.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Where

TP =True Positive

TN=True Negative

FP=False Positive

FN=False Negative

7. RESULTS AND DISCUSSION

The table below shows the values obtained by various performance evaluation parameters.

No.	Method of training	Learning rate	Momentum rate	Accuracy (%)	Precision	Recall
1	Simple Train & Test	0.7	0.6	35	0.5	0.35
2	Train after Association Rule Mining	0.9	0.8	62	0.76	0.62
3	K-Fold Cross validation of Train	0.2	0.3	91	0.95	0.91

In this table the considered parameters are method of training, learning rate, momentum rate, the accuracy precision and recall obtained on that specified learning and momentum rate. Learning and momentum rate that gives better training result is considered here. The values given to learning rate and momentum can range from 0.2 to 1.0.The three different training has three different results. So a better method selection is easier. Accuracy indicates how accurate the training method is. Here k-fold cross validation is better than MLP neural network training method.

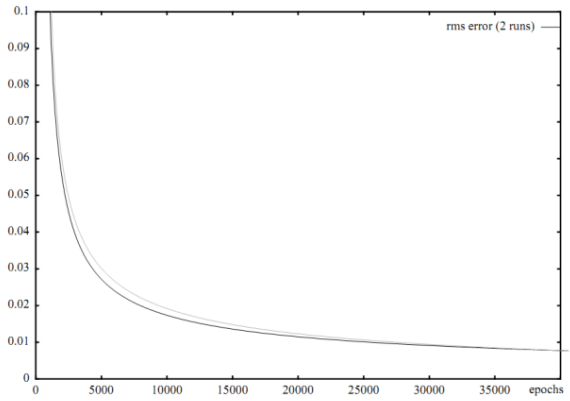

Figure 3.Genetic algorithm and back propagation algorithm

In this the result obtained by applying genetic algorithm is far better than the simple back propagation algorithm

8. CONCLUSION

This paper presented one use of data mining in the educational data mining field for predicting student's performance. Artificial neural network was used here for prediction.k-fold cross validation gives the most accurate result than basic training method and training after association rule mining. Association rule mining retrieved the most important attributes that affects the performance of student. Those attributes are mark of unit test, mark of assignment and attendance in the class.MLP training with this attributes gave a far better result than simple training.10 fold cross validation is used here for the training. The data set considered here is the real time dataset of marks of 300 students. As a future work fuzzy logic can be implemented to increase the performance evaluation result of the student.

9. REFERENCES

[1] Han J. and Kamber M.: "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000.

[2] Anwar, M. A., and Naseer Ahmed. Knowledge Mining in Supervised and Unsupervised ssessment Data of Students' Performance." 2011 2nd International Conference on Networking and Information Technology IPCSIT vol. Vol. 17. 2011.

[3]http://www.educationaldatamining.org/JEDM/index.php/JEDM

[4] V.O. Oladokun, Ph.D., A.T. Adebanjo, B.Sc., and O.E. Charles-Owaba, Ph.D. "Predicting Students' Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course."

[5] Refaat, M. Data Preparation for Data Mining Using SAS, Elsevier, 2007.

[6] S. M. Kamruzzaman and A. M. Jehad Sarkar
"A New Data Mining Scheme Using Artificial Neural Networks",2011.

[7] Amrender Kumar: "Artificial Neural Network for Data Mining".

[8] WEKA MANUAL

[9] Jung-Woo Ha." Classification using Weka (Brain, Computation, and Neural Learning)".

[10] Predicting Student Performance by Using Data Mining Methods for Classification; Dorina Kabakchieva Sofia University "St. Kl. Ohridski", Sofia 1000.

[11] Paresh Tanna, Dr. Yogesh Ghodasara:" Using Apriori with WEKA for Frequent Pattern Mining ".

[12]Baha Sen, Emine Ucar. Evaluating the achievements of computer engineering department of distance education students with data mining methods. Procedia Technology
1 262 – 267, 2012..

[13] http://www.theprojectspot.com/tutorial-post/creating-a-genetic-algorithm-for-beginners/3