International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue-1 E-ISSN: 2347-2693

Data Mining Based on Neural Networks for Education Data Forecasting

Kavita Pabreja

PhD., Birla Institute of Technology and Science, Pilani, Rajasthan, India
Associate Professor, Maharaja Surajmal Institute (GGSIP University) New Delhi, India
kavita_pabreja@rediffmail.com

www.ijcaonline.org

Received: Dec /26/2014 Revised: Jan/8/2015 Accepted: Jan/20/2015 Published: Jan/31/2015

***Abstract*—** Now-a-days, data mining has been used extensively in different domains of application for prediction. Data mining has demonstrated promising results in the field of educational prediction. Artificial Neural Networks in particular, find extensive application for understanding the peculiarities of education field but there is still a lot to be done as far as the Indian universities are concerned. In this paper, it has been verified that various personal and academic attributes of students can be used to predict the percentage of marks in graduation, using real data from the students of a Delhi state university's affiliates.

***Keywords*—**Educational Data Mining, Artificial Neural Network, Back Propagation, Academic Performance, Correlation analysis

I. INTRODUCTION

The application of neural networks in the data mining has become wider. Artificial Neural Network (ANN) which is designed to mimic the human brain, have complex structure and long training time but they have high acceptance ability for noisy data and have high accuracy. One of the key areas of the application of Education Data Mining is the development of student models that would predict student characteristics or performances in their educational institutions. Artificial neural network (ANN) has emerged during last decade as an analysis and forecasting tool in the field of education. In this study, the data mining based on neural networks has been used to forecast percentage of marks in graduation course. ANN has been trained based on student's class 10 and class 12 percentage, stream in class 12, attendance during graduation, time spent daily on study and social networking. The ANN hence trained has demonstrated promising results.

II. LITERATURE REVIEW

ANN has been applied in few such cases in the past. Related works can be found [1] that present a high level architecture and a case study for a prototype machine learning tool for automatically recognizing dropout-prone students and number of students who are likely to submit a written assignment (project) in university level distance learning classes. In a study by authors [2], algorithms are used which base their predictions on demographic data and a small number of project assignments. In another study [3], data of diabetes patients has been modeled and used it to predict the diabetes probability of any patient.

Corresponding Author: *Dr. Kavita Pabreja*
kavita_pabreja@rediffmail.com

III. PREPARE YOUR PAPER BEFORE STYLING

An ANN is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation [4]. 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. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. For the configuration, there are network functions used for training and testing of the network, as explained in following sections.

3.1 Network Function

The word '*network*'' refers to the inter–connections between the neurons in the different layers of each system. The most basic system has three layers. The first layer has input neurons which send data via synapses to the second layer of neurons and then via more synapses to the third layer of output neurons. More complex systems have more layers of neurons with some having increased layers of input neurons and output neurons. The synapses store parameters called

"weights" which are used to manipulate the data in the calculations.
The layers network through the mathematics of the system algorithms. The network function *f(x)* is defined as a composition of other functions $g_i(x)$, which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between variables, as shown in Fig. 1.

3.2 Training and testing the network

In an Artificial Neural Network, the system parameters are changed during operation, normally called the training phase. After the training phase, the Artificial Neural Network parameters are fixed and the system is deployed to solve the problem at hand (the testing phase). The Artificial Neural Network is built with a systematic step-by-step procedure to optimize a performance criterion or to follow some implicit internal constraint, which is commonly referred to as the learning rule [5]. The input/output training data are fundamental in neural network technology, because they convey the necessary information to "discover" the optimal operating point. The nonlinear nature of the neural network processing elements (PEs) provides the system with lots of flexibility to achieve practically any desired input/output map, i.e., some Artificial Neural Networks are universal mappers.

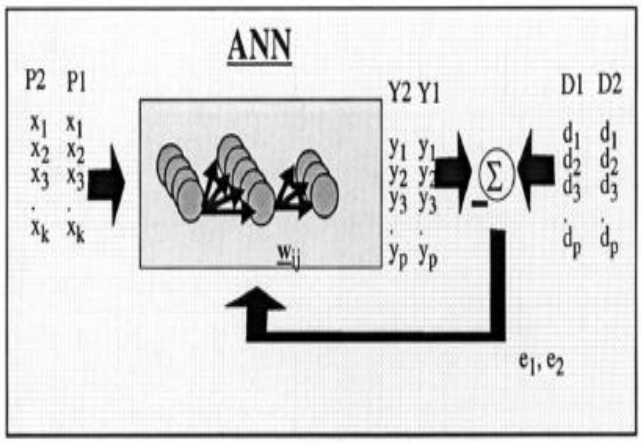

Fig.1. Style of neural computation

An input is presented to the neural network and a corresponding desired or target response set at the output (when this is the case the training is called supervised). An error is composed from the difference between the desired response and the system output. This error information is fed back to the system and adjusts the system parameters in a systematic fashion (the learning rule). The process is repeated until the performance is acceptable. It is clear from this description that the performance hinges heavily on the data. If one does not have data that cover a significant portion of the operating conditions or if they are noisy, then neural network technology is probably not the right solution. On the other hand, if there is plenty of data and the problem is poorly understood to derive an approximate model, then neural network technology is a good choice. In artificial neural networks, the designer chooses the network topology, the performance function, the learning rule, and the criterion to stop the training phase, but the system automatically adjusts the parameters. So, it is difficult to bring a priori information into the design, and when the system does not work properly it is also hard to incrementally refine the solution. But ANN-based solutions are extremely efficient in terms of development time and resources, and in many difficult problems artificial neural networks provide performance that is difficult to match with other technologies.

3.3 MLP Back Propagation Network

This is the most common neural network model, also known as supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using the historical data so that the model then can be used to produce the output when the desired output is unknown.
In this network, shown in Fig. 2, the input data are fed to input nodes and then they will pass to the hidden nodes after multiplying by a weight. This means that the artificial neurons are organized in layers, and send their signals “forward”, and then the errors are propagated backwards [6]. The idea of the backpropagation algorithm is to reduce this error, until the ANN learns the training data.

The network receives inputs by neurons in the input layer, and the output of the network is given by the neurons on an output layer. There may be one or more intermediate hidden layers which add up the weighted input received from the input nodes, associate it with the bias and then pass the result on through a nonlinear transfer function [5]. The output node does the same operation as that of a hidden layer. The output corresponding to the neuron model in Fig.2 is given by following equation:-

$$\mathbf{Y_k = \sum_{i=1}^{n} x_i\, w_i + b_k}$$

where x_i = input numbered i

w_i = weight for input i

b_k = bias

n = number of inputs

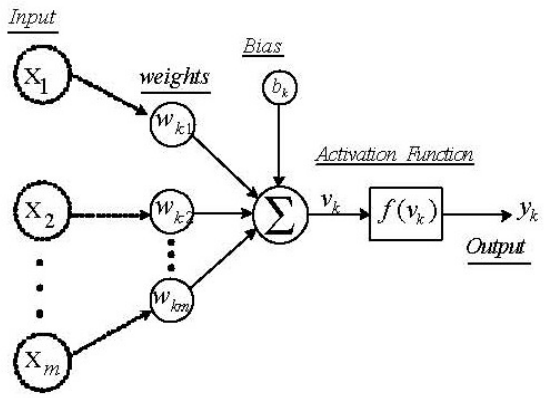

Fig.2. Neuron Model

IV CASE STUDY OF ACADEMIC RESULTS FORECASTING

4.1 Datasets used

The data has been collected as a part of the study under Major Project by students of UG programme. This study is based on real data collected from 1000 students of different colleges affiliated to a State University in Delhi. The students have provided the pre-university data i.e. percentage of marks in 10th and 12th Standard, Stream in 12th class. Other attributes used are as follows: daily Study Time (in Hrs), Social Networking Interests and Attendance in Current Semester. Based on these attributes, the academic performance in the university i.e. Percentage of marks in Graduation has been predicted. Data Cleaning has been done so as to fill up the missing values and noisy data with the most probable values.

3.2 Data Transformation: The numerical values have been either discretized or categorized. Given below is the table I depicting the data transformation done.

Variable	Description	Collected data	Transformed data
class 10	Percentage of marks obtained in class 10	Percentage of marks	No transformation
class 12	Percentage of marks obtained in class 12	Percentage of marks	No transformation
Stream in 12	Specialization in class 12	Science, Commerce, Humanities	No transformation
Attendance (Current Semester)	Percentage of students' presence in the class	Category of different percentages	Nominal data <40% = 1 40%-60% = 2 60%-75% = 3 >75% = 4
Study Time (in Hrs)	Daily study time in hours	Number of hours of study on daily basis	Nominal data 0-1hr = 1 1-2hr = 2 2-3hr = 3 3-4hr = 4
Social Networking	The sites used by student for social networking purpose	Facebook, Twitter, Whatsapp, wechat, line	Nominal Data Internet based (Facebook, Twitter) = 1 Mobile based (Whatsapp, wechat, line) = 2 Both =3 None = 4
Percentage in Grad	Percentage of marks in semesters I to V	Percentage of marks	No transformation

Table I Description of collected data

4.3 Technique used

ANN in this study was trained and simulated using Matlab 7.0 (matrix laboratory) designed and developed by Math Works Inc. For the training and testing of network, a two layer MLP Back Propagation network has been used. The input dataset comprises of class 10 percentage, class 12 percentage, Stream in class 12, Attendance (Current Semester), Daily Study Time (in Hrs), Social Networking interests. The output data corresponds to Percentage in Graduation. A sample of dataset is shown in table II. From this table, columns 1 to 6 are used as input and column 7 is used as target.

class 10	class 12	Stream in 12	Attendance (Current Semester)	Daily Study Time (in Hrs)	Social Networking	Percentage in Grad
58	64	2	1	3	4	68.5

69	67.2	2	1	2	4	78.93
53	83	2	3	2	4	70
80	65	1	1	1	4	81.71
76.7	69.2	1	3	1	4	80.3
78	62	1	4	1	4	78.8
67	74	2	1	4	3	63
61	58	1	2	4	3	72.6
73	70	2	2	4	3	67
61.8	61.25	2	2	4	3	73.4
86	73	1	3	4	3	85
86	73	1	3	4	3	85
80	78.6	1	3	4	3	87.4
58.5	70.25	2	3	4	3	72.65
63.6	69.8	2	3	4	3	81.6
79	56	1	4	4	3	70.18
81	69	1	4	4	3	74.4
75	55	1	4	4	3	75
78.4	77.5	1	4	4	3	76.77
66	63	2	4	4	3	75
63.8	78	2	4	4	3	83
40	75.25	3	4	4	3	63.65

Table II: Sample of Education data
(source: data collected as a part of study)

Before training, the inputs and outputs have been scaled so that they fall in the range[-1,1]. The following code has been used at Matlab prompt:-
[pn, minp, maxp, tn, mint, maxt] = premnmx (lnewin , lnewout);
The original network inputs and targets are given in the matrices lnewin and lnewout. The normalized inputs and targets, pn and tn, that are returned, will all fall in the interval [-1,1]. The vectors minp and maxp contain the minimum and maximum values of the original inputs, and the vectors mint and maxt contain the minimum and maximum values of the original targets.

4.4 Methodology

Different transfer functions for hidden and output layers were used to find the best ANN structure for this study. Transfer function used in hidden layer of the back propagation network is tangent-sigmoid while pure linear transfer function is used in output layer.
ANN developed for prediction of percentage of marks in graduation, is trained with different learning algorithms, learning rates, and number of neurons in its hidden layer. The aim is to create a network which gives an optimum result. The network was simulated using different Back propagation learning algorithms. The algorithm (trainlm) has been found to be giving best results.
trainlm is a network training function that updates weight and bias values according to Levenberg-Marquardt optimization. trainlm is often the fastest backpropagation algorithm in the Matlab toolbox

4.5 Results

The results of training and testing have been shown in Fig. 3. The mean square error is $0.5X10^{-2}$at 37 epochs which is quite satisfactory.

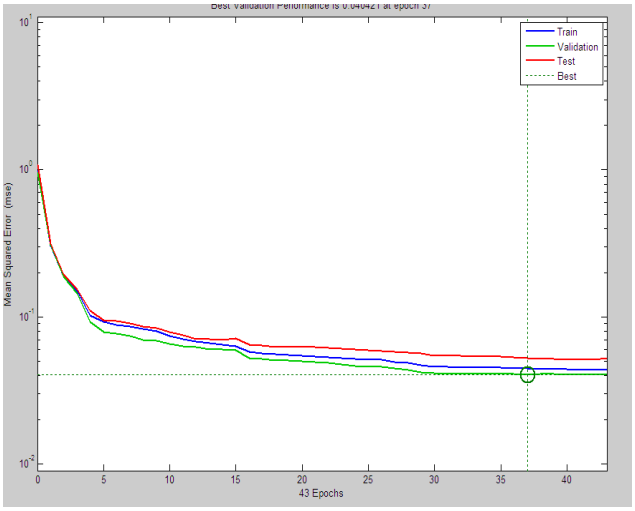

Fig. 3 Result of training, testing and validating ANN for education data using learning function trainlm

The correlation analysis where a correlation coefficient is reported representing the degree of linear association between actual percentage and the predicted percentage has been found as mentioned in table III and shown in Fig.4. A very strong correlation coefficient has been reported that advocates of the fact that the ANN with trainlm back propagation learning algorithm, transfer function namely tangent-sigmoid in hidden layer of the back propagation network and pure linear transfer function in output layer is able to predict the percentage in graduation quite correctly.

S.No.	Datasets	R (Correlation analysis)
1	Training	0.86419
2	Validation	0.87153
3	Testing	0.83766
4	All	0.86057

Table III – Details of values of R corresponding to different datasets under consideration

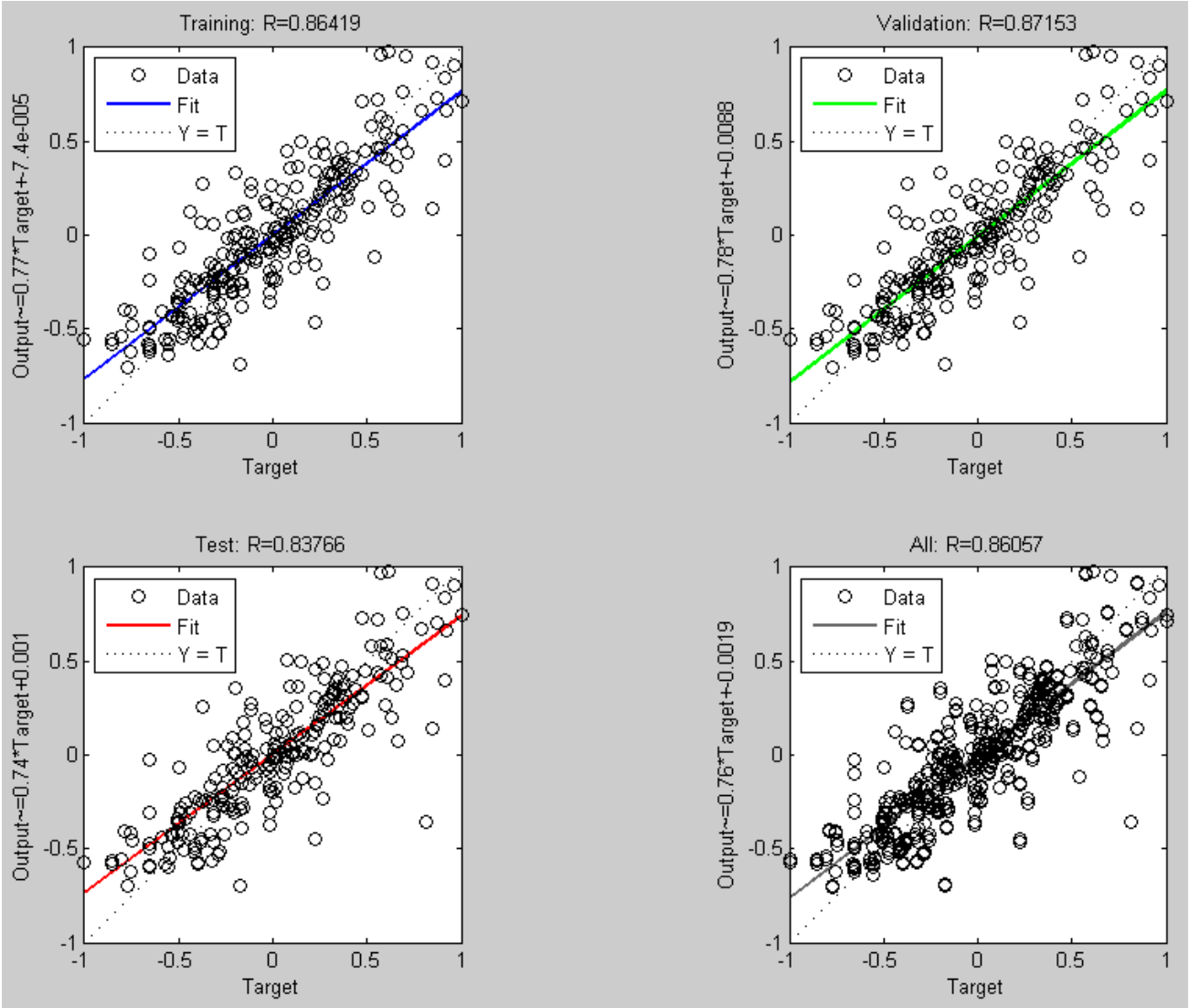

Fig. 4 Correlation analysis between Actual percentage and Predicted percentage

V. CONCLUSION

It is concluded that ANN has demonstrated promising results and is very suitable for solving the problem of education data forecasting. Using the input parameters as class 10 percentage, class 12 percentage, Stream in 12, Attendance (Current Semester), Daily Study Time (in Hrs), Social Networking interests, the ANN has been trained to predict percentage in graduation. This study has clearly brought out that Data Mining techniques when applied rigorously can help in providing advance information for forecast of graduation marks.

REFERENCES

[1] Kotsiantis S.B. and Pintelas P., A decision support prototype tool for predicting student performance in an ODL environment, International Journal of Interactive Technology and Smart Education, 1(4), p.p. 253-263, 2004.

[2] Kotsiantis S.B., Pierrakeas C. and Pintelas P., Predicting students' performance in distance learning using machine learning techniques, Journal of Applied Artificial Intelligence, 18(5), p.p. 411-426, 2004.

[3] Folorunsho O., Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database, International Journal of Advanced Research in Computer Science

and Software Engineering Research Paper, Volume 3, Issue 3, March 2013 ISSN: 2277 128X

[4] Sivanandam S.N., Sumathi S., Deepa S.N.(2009). Introduction to Neural Networks using Matlab, Tata McGraw Hill Education Private Ltd., 2009.

[5] Kosko B.(2005). Neural Networks and Fuzzy Systems, Prentice Hall of India Ltd., 2005.

AUTHORS PROFILE

Dr. Kavita Pabreja holds more than 19 years of experience with Educational institution and Industry. She is currently Associate Professor & Head - Computer Science, Maharaja Surajmal Institute, an affiliate of GGS Indraprastha University. She has teaching experience of over 14 years and she has also worked for more than 5 years with Indian as well as USA MNC. These companies include Rockwell International Overseas Corp., Parekh Microelectronics (I) Ltd., HCL Hewlett Packard Ltd. and Shyam Telecom Ltd.

She has completed her Ph.D. in the field of Data mining from BITS, Pilani. She has done M.S.(Software Systems), BITS, Pilani; AMIETE (eq. B.E. (Electronics and Telecommunication Engg.)) from IETE. She holds membership of many professional bodies *viz.* Senior Member of Computer Society of India, Member of Institute of Electronics and Telecommunication Engineers, Member of Indian Meteorological Society and Member of IACSIT, Singapore.

She has authored a book entitled "Application of Artificial Intelligence tools – Impact on Weather Prediction" with a renowned international publisher. She has designed and developed Workbooks and textbooks for the ICT Project, Punjab undertaken by Educational Consultants India Ltd. She has contributed twenty five papers in International Journals / Book/ International conferences. Her paper "Mapping of spatio-temporal relational databases onto a multidimensional data hypercube" presented at Einblick – Research Paper Competition held during Confluence 2010 organized by Amity University in association with EMC data storage systems (India) Pvt. Ltd. on January 22-23, 2010 was selected as the Best paper and awarded the FIRST prize.