IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

A Multilayered Back Propagation Algorithm to Predict Significant Attributes of UG Pursuing Students Absenteeism at Rural Educational Institution

S. Muthkumaran[1*], P. Geetha[2], E. Ramaraj[3]

[1]Dept. of Computer Science, Alagappa University, Karaikudi, India.
[2]Dept. of Computer Science, Dr. Umayal Ramanathan College for Women, Karaikudi, India
[3]Dept. of Computer Science, Alagappa University, Karaikudi, India

[*]Corresponding Author: muthumphil11@gmail.com

Available online at: www.ijcseonline.org

Accepted: 05/Dec/2018, Published: 31/Dec/2018

***Abstract*-**Recently Educational data mining has gained the attention of the researcher in the research industry and also in the society because of the availability of a large amount of data. There is a need for turning such data into useful information and knowledge. At present, there is a lack of well defined diagnostic algorithm to predict the reason for student absenteeism. It is critical to identify the most significant attributes in a dataset using the traditional statistical methods. This paper focuses on overcoming the difficulties involved in analyzing the student dataset by using Machine Learning Techniques. For mining purpose, Data pre-processing is done on the dataset which is a collection of questionnaire gathered from students in a semi-rural institution. Multilayered Back Propagation Algorithm was used to construct the neural network with weights and bias by applying a transfer function in the dataset. The highly influencing attributes having high weights and bias from the dataset was chosen and a Neural Network was constructed. This knowledge is used to identify the reason for the leave taken by the students and helps the management and staff members to improve the performance of the student.

Keywords: Educational Data Mining, Artificial Neural Networks, Multilayered Back Propagation Algorithm.

I. INTRODUCTION

The study of the biological nervous system gives rise to the development of new algorithm in Data mining which is called as Artificial Neural Network(ANN). ANN is a network of interconnected neurons, which works on the same principle of the working of a human brain[1]. The neural network algorithm uses a multilayer perceptron network consisting of three layers namely the input layer, the hidden layer, and the output layer. The ANN has a series of algorithms that try to identify underlying relationships in a set of data by using a process that works exactly the same as the working of the human brain. One of the main advantages of neural networks is the ability to changing input itself so the network produces the best possible result without redesigning the output parameters[2]. The major function of the neural network is to produce an output pattern for the given input pattern. The behavior of the neural network classification algorithm is as same as the human brain[3]. A neuron present in a neural network is a simple mathematical function which is used for capturing and organizing information according to a well-defined architecture.

This paper is organized as: Various related work done with Backpropagation Algorithm was discussed in Section II. The proposed methodology and the Architecture of the system Framework used to study the student's absenteeism were discussed in Section III. The Results obtained for the dataset using Back Propagation Algorithm is discussed in Section IV. The conclusion and Future works are discussed in Section V.

II. RELATED WORK

ParneetKaur, ManpreetSingh[4] presented a paper and in it, they identified students who have slow learning capacity in high school level. They compare various classification algorithms like Multilayer Perception, Naïve Bayes using WEKA which is an open source tool for data mining and compares the results produced from those above-mentioned algorithms. Cormac Reale, Kenneth Gavin[5] proposed a new Artificial Neural Network Algorithm to identify the variety of soils by using the geological background, particulate nature and engineering parameters used for building construction. Their method classifies 90% of the soil correctly and they give a better result than the various existing classification algorithms. Wai-Tak, Sheng-HsunHsu[6] proposed a new method for image retrieval from the search engine. Their method help to identify or search the image from search engine by without knowing the actual name of the image. The user who has a general idea of a particular image of who know only the desired of the image can retrieve image from a search engine using their proposed method. WeikuanJia, Dean Zhao[7] proposed a new method to overcome the difficulties present in using a Neural network such as choosing the number of hidden layers, setting the parameters for center and width of the hidden layers. Their proposed method gives the result in the form of a binary classification decision tree.

III. PROPOSED METHODOLOGY

This paper proposes a methodology to predict the highly influencing attributes in a student dataset for absenteeism. The following Figure1 illustrates the methodology to extract the significant attributes for student absenteeism. The first step deals with data collection using a questionnaire from a rural-based private institution situated at Villupuram District. Among 1500 students questionnaire are collected from 123 students were chosen on a random method. In it 85 students are male and 38 students are female. There are 30 attributes in the questionnaire from that 29 attributes excluding the name of the student were taken for study. Some of the attributes in the dataset are a job, assignment, college environment, sick, accident, test etc,

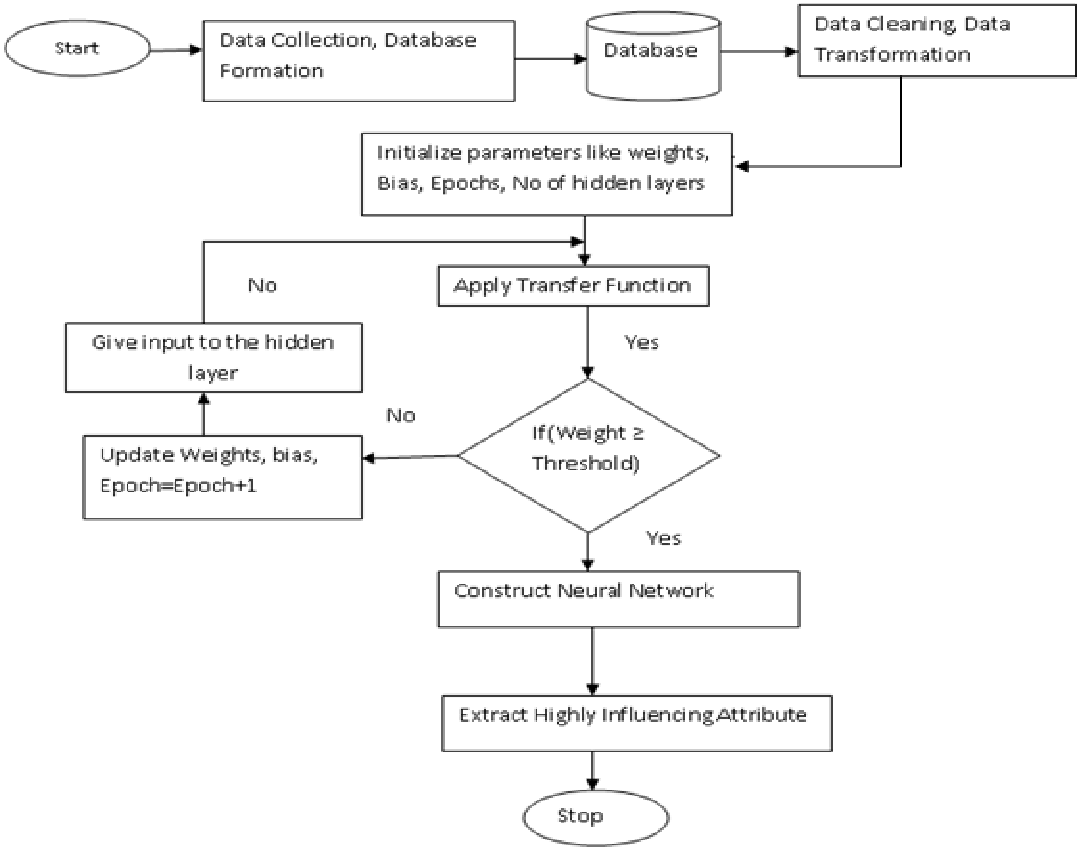
Start
Data Collection, Database Formation
Database
Data Cleaning, Data Transformation
Initialize parameters like weights, Bias, Epochs, No of hidden layers
No
Apply Transfer Function
Give input to the hidden layer
Yes
No
If(Weight ≥ Threshold)
Update Weights, bias, Epoch=Epoch+1
Yes
Construct Neural Network
Extract Highly Influencing Attribute
Stop

Figure 1: Architecture of the proposed model

The second step deals with Data preprocessing for mining purpose, the attributes which are left unanswered by the candidates were identified and filled using the relevant attributes which are answered by the remaining attributes of the student[8]. In data transformation method the attributes of the dataset which contains discrete attributes like Strongly agree, Agree, Neutral, Disagree, Strongly Disagree were converted into numerical attributes like 1 for Strongly agree, 2 for Agree, 3 for Neutral, 4 for Disagree, 5 for Strongly Disagree were put and the whole dataset is transformed into numerical dataset and stored in Excel 97-2003 format. The initial parameters for Multilayered Back Propagation algorithm like weight, bias, epochs, number of hidden layers were initialized[9]. The transfer function is applied to the dataset for converting the input parameter to the hidden layers and then to the output layer. If the weight calculated is greater than the threshold it is used for constructing the neural network otherwise its epoch value is incremented and given as the input to the hidden layer[10].

Algorithm: Backpropagation algorithm
Input: The training samples having attributes with numerical values, weights, bias, Epochs.
Output: A neural network with higher features attributes.
Step1: start the process
Step 2: Initialize the parameters like weights, bias, Epochs and select the number of hidden layers.
Step 3: For each input attribute in the dataset propagate the input from the input layer to the hidden layer using the transfer function $I_j=\sum_i w_{ij}\ O_i+O_j$.
Step 4: For each attribute in the hidden layer compute the weights using the activation function $O_j=\frac{1}{1+e^{-I_j}}$
Step 5: For each input attribute in the hidden layer from the first layer to the last layer compute the error with the next higher layer using $Errj=O_j(1-O_j)(T_j-O_j)$.
Step 6: Compute the weight for each attribute in the network using $\Delta Wij=(1)Errj\ O_i$ and update the weight using $W_{ij}=W_{ij}+D_{wi}$.
Step 7: For each input attribute in the network calculate bias using $\Delta\theta j=(1)$ Err and update the bias using $\theta j=\theta j+\Delta\theta j$.
Step 8: Repeat the process until the attributes with a higher threshold is reached
Step 9: Stop the process.

IV. RESULTS AND DISCUSSION

The dataset was implemented in Tanagra 1.4.50 which is an open source tool used for academic research purpose. The Back Propagation algorithm has the following parameters while implementing in Tanagra. The number of hidden layers used is 4, validation set proportion is 0.20, the learning rate is 0.15, Attribute transformation is set to none, Stopping rule is set to 100 maximum iterations, Error rate threshold is 0.0100. The values of each input attribute from the input layer to hidden layer are calculated using the transfer function $I_j=\sum_i w_{ij}\ O_i+O_j$ and the weight of each attribute is calculated using the activation function $O_j=\frac{1}{1+e^{-I_j}}$ and the values are listed in the following Table1.

Table 1: Weights from INPUT to HIDDEN layer

	Neuron "1"	Neuron "2"	Neuron "3"	Neuron "4"
Location of home	0.506	0.631	0.543	0.061
cinema	-1.140	-1.303	-1.280	-1.431
sick	-0.767	-0.837	-0.703	-0.739
Exam study	0.841	0.617	0.894	1.034
Job	-3.122	-2.982	-2.970	-3.798
Assignment	-1.146	-1.484	-1.150	-1.757
Fees	1.220	1.523	1.376	1.202
College environment	-1.365	-1.253	-1.285	-1.339
Staff problem	1.644	1.593	1.572	1.969
bias	-0.046	-0.031	-0.037	-0.088

The weights and bias are incremented for each step of iteration and the final values which come from the hidden layer to the output layer are listed in the following Table2.

Table 2: Weights From HIDDEN to OUTPUT layer

-	Male	Female
Neuron "1"	0.951	-0.951
Neuron "2"	1.384	-1.384
Neuron "3"	1.031	-1.032
Neuron "4"	1.582	-1.582
bias	-2.893	2.893

From the dataset 29 attributes are given as input to the Back Propagation Algorithm. The only 9 highly influencing attributes are identified using the proposed methodology and they are listed in the following Table 3.

Table3: Attribute contribution to Leave

Excluded attribute	Error rate	Difference	Statistics
Job	0.2846	0.065	1.7427
College environment	0.2602	0.0407	1.0892
Assignment	0.252	0.0325	0.8714
Fees	0.252	0.0325	0.8714
cinema	0.2439	0.0244	0.6535
Staff problem	0.2439	0.0244	0.6535
Exam study	0.2358	0.0163	0.4357
Location home	0.2276	0.0081	0.2178
sick	0.2276	0.0081	0.2178

From Table 3 it is clear that student put leave to college because of going to a job for earning money. The first reason for student's absenteeism is that they are going to a part-time job for earning money for their educational expenses. The second reason is the location of the college environment which is situated outside the town having lack of transport facilities make them put leave. The frequently conducted test in class and assignment were given to them make them put leave to the college. The release of new cinema is one of the major cause of student's absenteeism. When the semester exams are going to begin some students put leave to college and study for the exams at home. Students who are fallen sick put leave to college.

A. Performance Measures of the Backpropagation Algorithm

For every classifier, the performance of the classifier with regarding its classification of the dataset is measured with the help of the confusion matrix.

Table 4: Confusion Matrix for the dataset

Confusion matrix			
	Male	**Female**	**Sum**
Male	85(TP)	0(FN)	85
Female	27(FP)	11(TN)	38
Sum	112	11	123

For the student dataset, Tanagra has given the above confusion matrix which is listed in Table4 and this matrix is used to calculate the performance measures of this classifier. The overall performance measures of the Back Propagation Algorithm are listed in the table below.

Table 5: Performance measures

Accuracy	0.7804
Error rate	0.2195
Recall	1
Specificity	0.2894
Precision	0.7589
1- Precision	0.2411

The accuracy of a classifier is the percentage of tuples which are classified correctly and it is calculated using the formula.

Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}=\frac{85+11}{85+11+27+0}=\frac{96}{123}=0.7804.$

The tuples which are wrongly classified are called an error in machine learning. If the attributes in a data set are categorical then the misclassified tuples are expressed in error rate and it is calculated using the formula.

Error rate=1-Accuracy

Error rate=1- 0.7808=0.2195.

Recall or Sensitivity is the tuples which are correctly classified and it is calculated using the formula Recall=$\frac{TP}{TP+FN}=\frac{85}{85+0}=1.$

Specificity is a measure which is used to calculate the number of tuples which are wrongly classified and it is calculated using the formula.

Specificity=$\frac{TN}{TN+FP}=\frac{11}{11+27}=\frac{11}{38}=0.2894.$

Precision is a performance accuracy measure which gives the percentage of true positives (TP) when compared to the total number of tuples classified as positive events and it is calculated by the formula.

Precision=$\frac{TP}{TP+FP}=\frac{85}{85+27}=\frac{85}{112}$=0.7589.

1-Precision=1-0.7589=0.2411.

By using the results obtained from Tanagra for Back Propagation Algorithm which are listed in Table 1, Table 2, and Table 3. A Neural Network was constructed for the dataset and is given below in Figure 2.

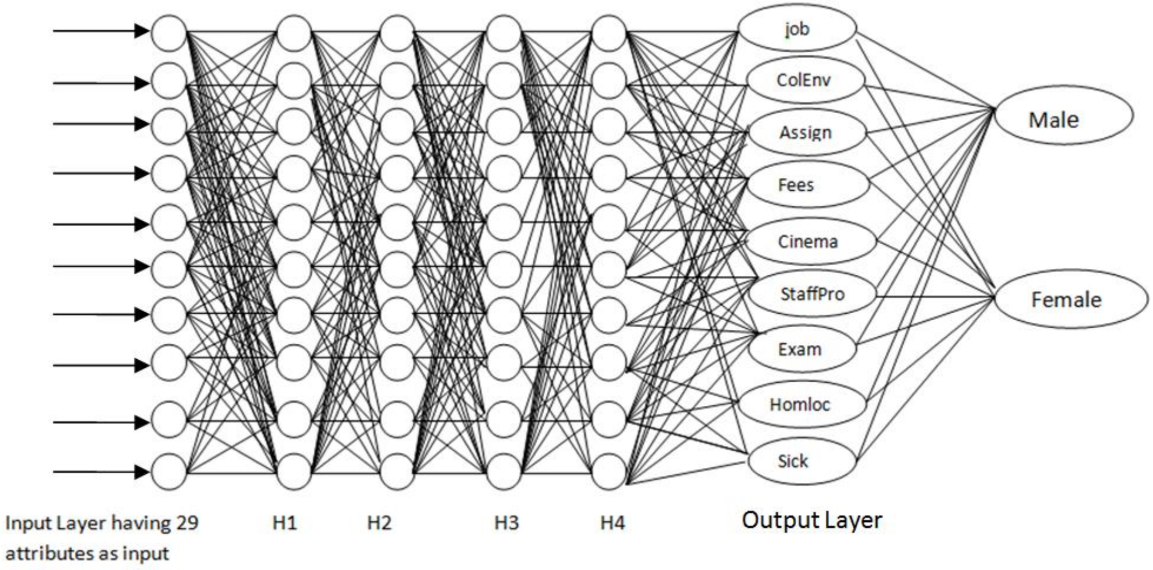

Figure 2: Neural Network Constructed for the dataset

V. CONCLUSION

This paper demonstrated that the Back Propagation Algorithm was used for prediction of Students Absenteeism and constructed an Artificial Neural Network using the dataset collected from students. The purpose of this work was to investigate and identify the reason for student absenteeism. From the results, it was found that attribute having more influence to take leave were identified using bias and weights of the Back Propagation Algorithm and the remaining attribute are removed from the dataset. The performance of the classifier was calculated using the performance measures of the Algorithm. These results act as guidelines for the management to develop the student's learning environment by discussing with the educational experts. Further work concentrates on applying Back Propagation Algorithm to various Educational domains.

REFERENCES

[1]. Gurpreet Kaur, Manreet Sohal, "IOT Survey: The Phase Changer in Healthcare Industry", International Journal of Scientific Research in Network Security and Communication, Vol-6, Issue-2, pp.34-39, 2016.

[2]. L.K. Ojha1 , L.K. Tiwary, R. Sharma," Information Communication Technology Integration in Education", International Journal of Scientific Research in Computer Science and Engineering, Vol.4, Issue.3, pp.14-15, June-2016.

[3]. Shubham Billus, Shivam Billus, Rishab Behl," Weather Prediction through Sliding Window Algorithm and Deep Learning", International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.5, pp.01-05, October -2018.

[4]. Kaur, P., Singh, M., &Josan, G. S. "Classification and prediction based data mining algorithms to predict slow learners in the education sector". Procedia Computer Science, Vol57, pp.500-508, 2015.

[5]. Reale, C., Gavin, K., Librić, L., &Jurić-Kaćunić, D. "Automatic classification of fine-grained soils using CPT measurements and Artificial Neural Networks". Advanced Engineering Informatics, Vol 36, pp.207-215, 2018.

[6]. Wong, W. T., & Hsu, S. H. "Application of SVM and ANN for image retrieval". European Journal of Operational Research, Vol173, Issue 3, pp.938-950, 2006.

[7]. Jia, W., Zhao, D., & Ding, L. "An optimized RBF neural network algorithm based on partial least squares and genetic algorithm for classification of small sample", Applied Soft Computing, Vol48, pp.373-384, 2016.

[8]. Bondarenko, A., Aleksejeva, L., Jumutc, V., &Borisov, A. "Classification Tree Extraction from Trained Artificial Neural Networks", Procedia Computer Science, Vol104, pp.556-563, 2017.

[9]. Gallego, A. J., Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R."Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation". Pattern Recognition, Vol74, pp.531-543, 2018.

[10]. Rizk, Y., Hajj, N., Mitri, N., &Awad, M. "Deep Belief Networks and Cortical Algorithms: A Comparative Study for Supervised Classification", Applied Computing and Informatics, 2018.

[11]. Ougiaroglou, S., Diamantaras, K. I., &Evangelidis, G."Exploring the effect of data reduction on Neural Network and Support Vector Machine classification", Neurocomputing, 2018.

[12]. Zhang, J., Williams, S. O., & Wang, H. "Intelligent computing system based on pattern recognition and data mining algorithms", Sustainable Computing: Informatics and Systems, 2017.

Author's Profile

S.Muthukumaran is a Research Scholar in the Department of Computer Science, Alagappa University, Karaikudi. His main Research Interest includes Data Mining, Big Data Analytics, and Internet of Things. He has published 8 International Journals.

P.Geetha is working as an Associate Professor in the Department of Computer Science at Dr. Umayal Ramanathan College for women, Karaikudi. Her Research Interest includes Data Mining, Network Security and Internet of Things. She has published 13 International Journals.

E. Ramaraj is working as the Professor and Head of the Department of Computer Science, Alagappa University, Karaikudi. He has the sound knowledge in many research fields especially in Data Mining, Network Security, Remote Sensing, and Big Data & Analytics. He has published more than 100 international journals.