VISUAL ANALYSIS OF EDUCATIONAL DATA USING NEURAL NETWORK BASED CLUSTERING AND CLASSIFICATION APPROACH

PRATIYUSH GULERIA AND MANU SOOD
Department of Computer Science, Himachal Pradesh University, Shimla
Himachal Pradesh, India
pratiyushguleria@gmail.com, soodm_67@yahoo.com

ABSTRACT: To increase the quality of education and to find solution to problems arising from complex educational dataset and competitive environment among the academic institutions, Educational Data Mining is receiving great attention. Student's performance is of great concern to the higher education. In this paper, we have applied two approaches for educational data mining. The first approach is based on self-organizing map (SOM) which is a type of ANN (Artificial neural network) that is trained using unsupervised learning to produce low-dimensional views of high-dimensional data. Using this approach, we have clustered students based on certain attributes into natural classes so that similar classes are grouped together. The second approach uses pattern recognition through two-layer feed-forward network to classify inputs into a set of target categories.

KEYWORDS: ANN, Classification, Clustering, Mining, Neural, Pattern Recognition, Self-organizing map.

INTRODUCTION

Data Mining is playing significant role in educational systems where education is the key input for social development [1]. The main challenge of Institutions is to deeply analyze their performance in terms of student performance, teaching skills and academic activities. There are many Data Mining Techniques like K-means Clustering, Decision trees, neural networks, Nearest Neighbor; Naive Bayes etc are being used in Educational Data Mining. Using these methods, knowledge can be extracted such as classification, clustering and association rules which are helpful in analyzing quality of education [2]. Classification is a supervised learning based DM technique where training data set is input for the classifier [3] whereas Clustering is the unsupervised classification of patterns into clusters [4].

Neural Networks is basically a group of interconnected neurons that use computational or mathematical models and processes information. Neural Network usually learns by examples. It consists of three layers i.e. input, output and hidden. Each node from input layer is connected to a node from hidden layer and every node from hidden layer is connected to a node in output layer and usually some weights are associated with every connection [5].

A self-organizing map consists of neurons and each neuron is associated with weight vector of the same dimension as the input data vectors.SOM is an unsupervised neural network algorithm and projects high-dimensional data onto a two-dimensional map. In this similar data items are mapped to nearby locations. Using Data Clustering, we extract previously unknown and hidden pattern from large data sets [6]. A two-layer feed forward network is used for pattern recognition which

transforms sets of input signals into set of output signals. This type of learning is unsupervised and there is good internal representation of the input [7].

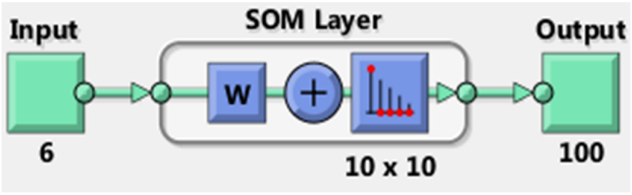

Figure 1. Self-Organizing Map Network

Fig.1 shows the process of self-organizing map clustering which consists of a competitive layer that classifies Educational Dataset with any number of dimensions into as many classes as the layer has neurons. Educational Dataset consists of attributes mentioned in Table I to predict their class result using classification techniques. After collection of Data, there is pre-processing of Data in which data is cleaned and transformed into an appropriate format to be mined.

LITERATURE REVIEW

In [8] has discussed about Cluster Analysis and advanced method of Neural Network based on Kohonen self-organizing map. In [9], author presents pattern classification method which is self-organizing and produces fuzzy outputs.Multi-layer Perceptron neural network with Back-propagation algorithm and Self-organizing(Kohonen's) maps are used for classifying data [10] and solving pattern recognition tasks.

In [11] author suggested neural network based approach of mining classification rules from given databases in three phases i.e. constructing and training a network, then pruning the network while maintaining classification accuracy and extraction of symbolic rules from the pruned network. Inductive Modelling Technique applied on dataset results in extraction of six features which are used as input to self-organizing map [12].Data Mining has played an important role in Education and according to author in [13], Students performance can be improved using Data Clustering and Neural Networks. Artificial intelligence has enabled the development of more efficient student models which represent and detect a broader range of student behavior. Authors in [14] have proposed the application of an artificial neural network for predicting student's performance in final exams. Kohonen Self-organizing feature map properties can be used within the data mining and knowledge discovery process. James Malone et al. [15] proposed a technique for the automatic extraction of rules from trained SOMs which are responsible for clustering. Self-Organizing Maps helps not only in data-clustering but also to visualize multi-dimensional data [16]. Marie Khair et al. [17] Proposed a decision-making procedure for both students and advisors using Educational Data Mining Techniques and Neural Networks.

PROBLEM DEFINITION

In this paper, problem formulation is related to pattern recognition and clustering. In clustering main problem is not only to group samples into classes based on the similarity between samples but also to classify unknown inputs accordingly whereas problem with pattern recognition is that

inputs are associated with different classes and complex decision boundary problems over many variables.

Therefore, we have proposed a neural network that cluster students based on their academic attributes and classify the student's educational data set by minimizing cross-entropy results and low misclassifications. It also uses the attributes of neighbourhoods to train the network to produce the correct target classes.

DATA MINING PROCESS

Data Preparations

We have initially collected the data set of 120 Graduate Students from one of the educational institutions.

Data Selection

Now analyse the class result of these Students taking 5 Attributes of student record i.e. Class Performance, Attendance,Assignment,Lab Work, Sessional Performance (Aggregate of Ist and IInd Sessional). The student's class result is predicted after analysing performance in these attributes**.** Attributes and Educational Data Set of students is given in Table 1 and Table2.

Table 1. Attributes

Attributes	Description	Coding
ClassID	Roll No of Student	Numeric Value
CLP	Class Performance of the student	{Excellent= "Above 80",Good = "70-80",Average= "60-70",Poor= "Below 60"}
Sessional	Aggregate of Ist and 2nd Sessional	{Good= "75-100",Average= "60-75",Poor= "Below 60"}
ASSGN	Assignments	"A"=8-10, "B"= 6-7,"C"=1-5}
ATTD	Attendance	{Good = ">=90%",Average = ">=75% or above" ,Poor = "below 75%"}
LW	Lab Work	{Good= "20-25",Average = "15-20",Poor= "Below 15"}
CLR	Class Result of Students	{First= "Above 70%",Second = "60-70%",Third= "50-60",Fail= "Below 50"}

Clustering and Pattern Recognition using Neural Network

Data Pre-processing can be simple transformations performed on single variables.SOM algorithm uses Euclidean metric to measure distances between vectors [18].ANN consists of training and learning phases. In training phase, with some input there is predictable output, provided certain pattern of weights. Here, Neural Network does not require a step-by-step procedure to perform desired task instead the network can be taught to do the task. As the training process proceeds, the weights will converge to values and perform some useful computations knowing nothing initially and moves on to gain knowledge. This phase is known as unsupervised learning phase [19].

Steps followed in self-organizing map clustering shown in Table 3

Distances between neurons are calculated from their positions with a distance function. Using dist function, distance from a particular neuron to its neighbours is calculated. The dist function calculates the Euclidean distance from a home neuron to any other neuron. There are some

Table 2. Educational Data Set

CLASSID	CLP	SESSIONAL	ATTD	ASSGN	LW	CLR
1	78	79	89	9	23	79
2	79	80	89	10	22	81
3	71	70	76	10	23	71
4	52	79	55	4	10	80
5	52	50	45	3	12	51
-	--------	----------	----------	----------	----------	--------

command-line functions shown in Table 4 that can be used to analyse the resulting clusters once the network has been trained.

Table 3. SOM Clustering Steps

Step1	Random input is selected.
Step2	Winner Neuron is computed.
Step3	Neurons are updated.
Step4	Repeat for all input data.
Step 5	Each weight vector then moves to the average position of all the input vectors for which it is a winner or for which it is in the neighbourhood of a winner.
Step 6	Finally classification of input data is there.

Classification using pattern recognition

In pattern recognition, neural network classify inputs into a set of target categories. Neural Pattern Recognition helps to select data, create and train a network, and evaluate its performance using cross-entropy and confusion matrices.Fig.2. shows a two-layer feed-forward network, with sigmoid hidden and output neurons which classify vectors arbitrarily well, given enough neurons in its hidden layer.

In pattern recognition using Neural Network, input dataset is presented to the network and target dataset define in Table 5 shows the desired network output.

Table 4. Command-Line Functions

Command-line Function	Description
plotsomtop(net)	Plots the self-organising map topology.
plotsomd(net)	Plots self-organizing map neighbour distance and shows how close each neuron's weight vector is to its neighbours.
plotsomhits(net, inputs)	Plots self-organizing sample hits. Each neuron shows the number of input vectors that it classifies.
plotsomplanes(net)	Plots self-organizing map weight planes and shows a weight plane for each of the input features.
plotsomnc(net)	Plots self-organizing map neighbour connections.

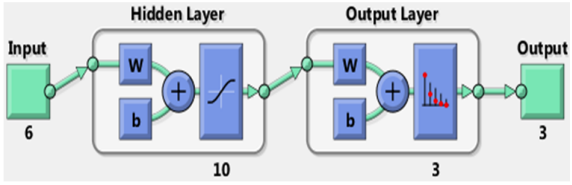

Figure 2. Two-Layer Feed-Forward Network

Table 5. Desired Network Output

1	1	1	1	1	0	0	0	0	0	0	-	-
0	0	0	0	0	1	1	1	1	1	0	-	-
0	0	0	0	0	0	0	0	0	0	1	-	-

Network is trained and we classify the inputs according to the targets. Training multiple times will generate different results due to different initial conditions and sampling. After training, it shows results in the form of Cross-Entropy and Percent Error. Minimizing Cross-Entropy results in good classification. Lower values are better and Zero means no error. Percent Error indicates the fraction of samples which are misclassified. A value of 0 means no misclassifications, 100 indicates maximum misclassifications.

Another measure of how well the neural network has fit the data is the confusion plot. Here the confusion matrix is plotted across all samples.

RESULTS AND DISCUSSIONS

In this paper, clustering and pattern recognition applied on the student's training dataset mentioned in Table II is implemented with the help of MATLAB simulink.

For clustering, the dataset is loaded into Mat lab and normalized. After the dataset is ready, Self-Organizing Map is trained. Inputs 'data' is a 120x6 matrix, representing static data consisting of 6 samples of 120 elements. SOM neural network cluster students data into classes topologically based on their performance. After training the network, it shows following plots: SOM Neighbour Distances, SOM Weight Planes, and SOM Sample Hits.

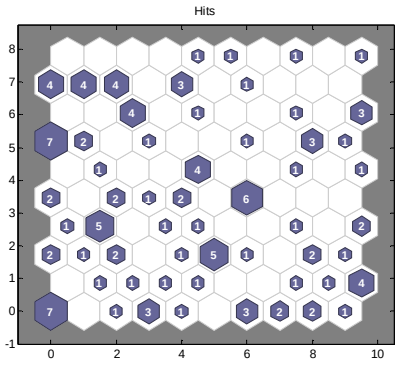

Figure 3. SOM Sample Hits

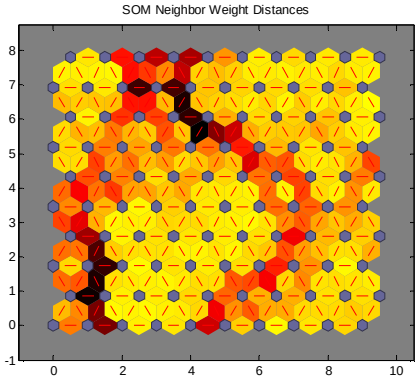

Figure 4. SOM Neighbour Distances

Fig.3 shows the calculation of the classes for each student attribute and the number of students belonging to that class. Areas of neurons with large numbers of hits indicate classes representing similar and highly populated regions of the feature space whereas areas with few hits indicate sparsely populated regions of the feature space.

Fig.4 shows how distant in terms of Euclidian distance, each Neuron's class is from its neighbours. Connections which are bright indicate highly connected areas of the input space. While dark connections indicate classes representing regions of the feature space which are far apart, with few or no students between them. Long borders of dark connections separating large regions of the input space indicate that the classes on either side of the border represent students with very different features.

Fig.5 shows a weight plane for each of the six input features. They are visualizations of the weights that connect each input to each of the 100 neurons in the 10x10 hexagonal grids. Darker colors represent larger weights. If inputs have similar weight planes, their colour gradients may be the same or in reverse it indicates they are highly correlated.

For pattern recognition, we input the dataset shown in Table 2 and desired target dataset in Table 5 and it randomly divide up the 120 samples into 3 types shown in Table 6. Input is a 6x120 matrix, representing static data: 120 samples of 6 elements. Target is a 3x120 matrix, representing static data: 120 samples of 3 elements.

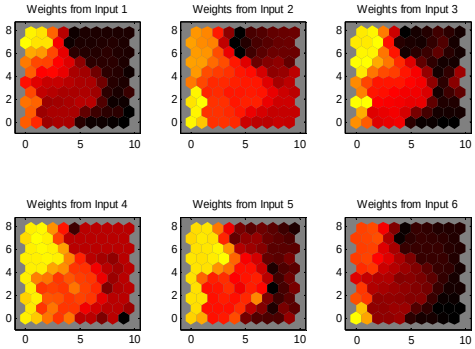

Figure 5. SOM Weight Planes

Table 6. Samples for Validation and Testing

Training	These are presented to the network during training, and the network is adjusted according to its error.	70%	84 samples
Validation	These are used to measure network generalization, and to halt training when generalization stops improving.	15%	18 samples
Testing	These have no effect on training and so provide an independent measure of network performance during and after training	15%	18 samples

Fig.6.shows the best validation performance at epoch 27 after the network is trained to classify the inputs according to the targets and Fig.7.shows Confusion Matrix plotted across all samples with zero misclassifications. The confusion matrix shows the percentages of correct and incorrect classifications. Correct classifications are the green squares on the matrices diagonal whereas incorrect classifications form the red squares. If the network has learned to classify properly, the percentages in the red squares should be very small, indicating few misclassifications.

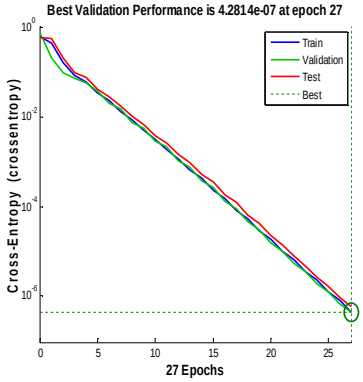

Figure 6. Training Performance

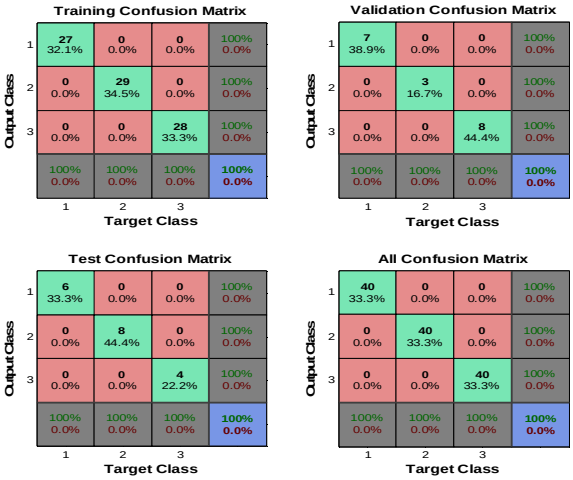

Figure 7. Confusion Matrix

CONCLUSION

In this paper, Using Self-Organizing Map Neural Network Clustering and pattern recognition data mining is achieved by partitioning data into related subsets and clustered the students based on attributes like Class Performance, sessionals and attendance in class. Visual Results shows the self-organizing maps topology of neurons where each neuron has learned to represent a different class of students with adjacent neurons typically representing similar classes.

This study helps in clustering students according to their academic performance as well as pattern recognition and regularities in data enhances the decision-making approach to categorize the students and monitor the performance of those students who are showing poor academic performance. Results are achieved with minimum cross-entropy and percent error which indicates good classification and minimum misclassifications of sample.

REFERENCES

Sonali Agarwal, G. N. Pandey, and M. D. Tiwari, “Data Mining in Education: Data Classification and Decision Tree Approach”, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012.

Alaa El-Halees, “Mining Students Data to Analyse Learning Behaviour: A Case Study”, Available online at: https://uqu.edu.sa/files2/tiny_mce/plugins/filemanager/files/30/papers/f158.pdf.

Mohd. Mahmood Ali, Mohd. S. Qaseem,Lakshmi Rajamani, A. Govardhan, “Extracting useful rules through improved decision tree induction using information entropy”, International Journal of Information Sciences and Techniques (IJIST),Vol.3, No.1, January 2013.

A.K. Jain, M.N. Murty, P.J. Flynn, “Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999.

M. Cilimkovic, “Neural Networks and Back Propagation Algorithm”, Institute of Technology Blanchard town, Blanchard town Road North Dublin 15, Ireland.

Md. Hedayetul Islam Shovon,Mahfuza Haque,"An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree", International Journal of Advanced Computer Science and Applications,Vol.3, No. 8, 2012.

Teuvo Kohonen,"The Self-Organizing Map", Proceedings of the IEEE, vol 78, No.9, September 1990.

P.Dostal, P.Pokorny, "Cluster Analysis and Neural Network", 2009.

D.T.Pham, E.J.Bayro-Corrochano,"Self-organizing Neural-Network-Based Pattern Clustering Method with Fuzzy outputs", Pattern Recognition, Vol.27.No.8, pp.1103-1110, 1994.

Jiri Stastny, Pavel Turcinek, Arnost Motycka, "Using Neural Networks for Marketing Research Data Classification", Mathematical Methods and Techniques in Engineering and Environmental Science, ISBN: 978-1-61804-046-6.

Hongjun Lu, Rudy Setiono, Huan Liu, "Effective Data Mining Using Neural Networks", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, December 1996.

D. Novak, P. Kordık, M. Macas, M.Vyhnalek, R. Brzezny, L. Lhotska, "School Children Dyslexia Analysis using Self Organizing Maps", 2004 Proceedings of the 26th Annual Embs International Conference, September 3-5, San Francisco, USA.

Chady El Moucary, Marie Khair, Walid Zakhem, "Improving Student's Performance Using Data Clustering and Neural Networks in Foreign-Language Based Higher Education", The Research Bulletin of Jordan ACM , Vol.II (I II), ISSN: 2078-7952 (print); 2078-7960 (online).

Ioannis E.Livieris, Konstantina Drakopoulou, Panagiotis Pintelas, "Predicting students' performance using artificial neural networks".

James Malone, Kenneth McGarry, Stefan Wermter and Chris Bowerman, "Data Mining using Rule Extraction from Kohonen Self-Organizing Maps".

Pavel Stefanovic, Olga Kurasova, "Visual analysis of self-organizing maps, Nonlinear Analysis: Modelling and Control", 2011, Vol. 16, No. 4, 488–504.

Marie Khair, Walid Zakhem ,Chady El Moucary, "Solving Probation and Change-of-Major Issues in Higher Education Using Data Mining Techniques", International Journal of Multidisciplinary Sciences and Engineering, Vol. 3, No. 11, November 2012.

Juha Vesanto, Johan Himberg, Esa Alhoniemi and Juha Parhankangas, "Self-organizing map in Matlab: the SOM Toolbox", Proceedings of the Matlab DSP Conference 1999, Espoo, Finland, November 16-17, pp.35-40, 1999.

Jing Li, "Information Visualization with Self-Organizing Maps", Available online at: http://www14.in.tum.de/konferenzen/Jass05/courses/6/Papers/09.pdf