VOL. 11, NO. 9, MAY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A NOVEL NEURAL NETWORK APPROACH TO DATA CLASSIFICATION

K. G. Nandha Kumar[1] and T. Christopher[2]
[1]PG and Research Department of Computer Science, Govt. Arts College, Udumalpet, Tamilnadu, India
[2]PG and Research Department of Computer Science, Govt. Arts College, Coimbatore, India
E-Mail: kgnandhakumar@gmail.com

ABSTRACT

Data classification is a major task in data mining paradigm. In this paper an artificial neural network approach is proposed for data classification. In this approach data classification is accomplished through a cluster analysis. It is a two-pass process, clusters are created in the first step and classification is achieved from the results of first pass. A self organizing map neural network (SOMNN) is used for clustering in the first pass. In the second pass classification task is completed by using multilayer neural networks (MNN). Basically SOM is an unsupervised neural network and multilayer networks are supervised neural network, hence this approach is a hybrid method. Nine hybrid neural networks (HNN1 to HNN9) are constructed from the combination of above said methods and are experimented. Performance of each hybrid neural network is evaluated by using metrics such as accuracy, precision, recall, and F-measure. Feedback of library users is used as data set for classification.

Keywords: data classification, artificial neural networks, self organizing map, hybrid neural network.

INTRODUCTION

In recent decades many data analysis methods, techniques and tools have been invented by researchers of different functional fields, among them data mining deals with knowledge extraction from raw data. Data classification is a data mining task which categorizes the large amount of data under known labels. This is suitable to analyze survey data of academic libraries. To achieve better data classification there are hundreds of techniques including statistics, traditional algorithms and machine learning algorithms available. In the last few decades machine learning methods have gained more attention among the research communities and they are in vast range of fields. Artificial neural networks (ANN) are one of the machine learning methods. In data analysis researches ANNs are applied equally with regular statistical methods. They are proven in terms of accuracy and reliability. ANN techniques are inspired by biological neurons of human brain and one of the machine learning techniques. They are applied in various fields such as data analytics, engineering, finance and business and so on. Such type of computational intelligence methods are proven in terms of reliability, accuracy, and predictability.

Here a novel method is proposed and tested. The proposed method is an artificial neural network (ANN) based data mining technique. Data mining is a set of data analysis methods which discovers unknown facts from raw data. They are used to extract knowledge from large data sets. Data classification, data clustering and mining associations are primary data mining tasks. Data classification task is accomplished by artificial neural network technique. The purpose of this proposed method is to obtain better classification results in terms of accuracy and other performance metrics.

LITERATURE REVIEW

Many researchers have applied data mining techniques to analyze different kinds of library data with several algorithms, methods and tools. Ping YU (2011) has developed a library reader management model to identify characteristics of readers' book borrowing in order to evaluate individual service. He has used book lending history as data and applied cluster analysis as data mining technique. Tingting Zhu and Lili Zhang (2011) have applied decision tree 64.5 algorithms to analyse the needs of users of university library. The data of Changchun institute of technology have been used as input. It is a classification analysis and by applying it various needs of users are classified.

Runhua Wang, Guoquan Liu, Yi Tang, and Yan Li (2011) have applied K-means clustering algorithm to analyze readers' characteristics. Books circulation records and master list of books are used as data. Through the analysis they have found that the library should have added more books and they recommend the authorities to do so base on results of analysis. Runhua Wang, Yi Tang and Lei Li (2011) have applied back propagation neural network to predict library circulation of future. They have constructed three layered neural network for prediction process. It has been used to improve library management services.

Suresh Kumar PK (2012) has analysed user satisfaction and quality of service of the university libraries in Kerala by using RATER analysis based on seventeen variables. A neural network based assessment method is developed and applied by Sun Yanhua (2012) to assess college library website. The results are used for further decision making purpose. Keita Tsuji *et al*., (2012) have applied association rule mining technique to analyze circulation records of the library. Association rule mining is a data analysis technique which finds relations, and dependencies among activities or variables. As the result of this study they have developed book recommendation model for the library which will suggest what kind of books shall be added to the library.

Veepu Uppal and Gunjan Chindwani (2013) have done a literature survey on applications of data mining techniques in library system. They have listed out that how

the three core techniques of data mining namely classification, clustering and association rule mining can be applied on library data. Mohd Dasuki Sahak and Mohamad Noorman Masrek (2013) have done a comparative analysis on usage of first and third year students of medical and health sciences of University Putra Malaysia. The information seeking behaviour of readers is analyzed by using traditional statistical methods.

Hara Brindesi, Maria Monopoli and Sarantos Kapidakis (2013) have investigated the information seeking behaviour and searching habits of Greek physicists and astronomers. The study was done at National university and Kapodistrain universities of Athens. The data were collected from undergraduate students of departments of astrophysics, astronomy and mechanics of the above said universities. Their method was regular statistical techniques. Pareek AK and Madan S. Rana (2013) have applied statistical methods to study information seeking behaviour and to analyze library use pattern of researchers at the Banasthali University of India. They have studied various components; such as frequency of visit, purpose of visit, methods of information seeking, use of library sources and services, purpose of information seeking, problems faced in using library, library rating and use of e-resources. Pijitra Jomsri (2014) has applied association rule mining technique to develop a book recommendation for digital libraries. His study was based on user profiles. He found various associations in order to recommend new books. His technique was entirely data mining based.

PROPOSED METHOD AND EXPERIMENTS

A two-pass hybrid method is proposed for data classification as follows.

First pass (kohonen self organizing maps)

1. Initializing the weights and setting topological parameters.
2. Computing the square of Euclidean distance as follows for each input vector.

$$D(j) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - w_{ij})^2$$

3. Finding the final unit index J, so that D(J) is minimum.
4. Updating weights for all j as follows.

$$w_{ij}(new) = (1 - \alpha)\, w_{ij}(old) + \alpha\, x_i$$

5. Updating the learning rate α by using α(t+1)=0.5α (t).
6. Reducing topological parameter.
7. Testing for stopping condition.

Second pass (multilayer neural networks)

1. Initializing the weights and bias.
2. Calculating net input as follows.

$$y_{in} = b + \sum_{i=1}^{n} x_i w_i$$

3. Calculating output by applying activation function over the net input as follows.

$$y = f(y_{in}) = \begin{cases} 1 \; if \; y_{in} > \theta \\ 0 \; if \; -\theta \le y_{in} \le \theta \\ -1 \; if \; y_{in} < -\theta \end{cases}$$

4. Updating weight and bias.
5. Testing for stopping condition.

Cluster analysis is a technique to categorize raw data into different classes based on similarities in properties. It is done by using Kohonen-SOM neural networks. Classification is a technique to assign values under certain named classes and it is accomplished by using MNN. Here classification through clustering approach is adopted. Since it is a machine learning method, training is important for a newly built neural network. After certain amount of training is given the network will learn and work accordingly. So one third amount of data set is used as training data and then remaining data set is processed by the neural network. Students' feedback data is used for this classification task. It has 2467 records with 8 attributes and 2 classes.

Network architecture plays significant role in any type of neural network. Number of neurons in each layer, number of layers and connectivity style are parameters of the architecture. To test this two-pass method, nine different networks were built for second pass based on the number of neurons in the processing layers. By combining these nine networks of second pass with SOM of first pass, nine different hybrid neural networks HNN1, HNN2, HNN3, HNN4, HNN5, HNN6, HNN7, HNN8, and HNN9 were built. Number of neurons of input layer, hidden layer and output layer of each hybrid network is represented in the Table-1.

Table-1. Number of neurons in input, hidden, and output layers.

ANN Type (SOM+MNN)	HNN1	HNN2	HNN3	HNN4	HNN5	HNN6	HNN7	HNN8	HNN9
No. of neurons	8-8-1	8-10-1	8-12-1	8-14-1	8-16-1	8-18-1	8-20-1	8-22-1	8-24-1

RESULTS AND DISCUSSION

The performance of all the neural networks are evaluated by using evaluation metrics; accuracy, precision, recall, and F-measure. The performance of all the networks in terms of metrics is shown in the Table-2. The formulas are:

- Accuracy = (tp+tn)/(tp+fp+tn+fn)
- Precision (P) = tp/(tp+fp)
- Recall (R) = tp/(tp+tn)
- F-Measure = (2*P*R)/(P+R)
- tp = true positive classification & tn = true negative classification
- fp = false positive classification & fn = false negative classification

Table-2. Performance of hybrid neural networks.

ANN Type / Performance metrics	HNN1	HNN2	HNN3	HNN4	HNN5	HNN6	HNN7	HNN8	HNN9
Accuracy %	76.21	76.23	76.31	76.79	76.91	76.92	**77.14**	77.10	77.11
Precision	0.647	0.647	0.683	0.731	0.712	0.754	0.768	**0.772**	0.770
Recall	0.617	0.639	0.673	0.735	0.728	0.749	**0.759**	0.720	0.698
F-measure	0.632	0.643	0.678	0.733	0.720	0.751	**0.763**	0.745	0.732

While examining the performance of the nine neural networks, following facts are found. There is no significant change in the accuracy of HNN1, HNN2, and HNN3. The number of hidden layer neurons of these networks produces similar accuracy but the scores of precision, recall and F-measure of HNN3 are greater than HNN1 and HNN2. In the first three networks set, HNN3 shows considerable performance. When compare the performance of neural networks HNN4, HNN5, and HNN6 of second set, the HNN6 shows better accuracy and scores than other two networks. In the third set of neural networks, HNN7 produces better classification results in terms of accuracy and also it shows better scores in terms of recall and F-measure. HNN8 provides better results in terms of precision. HNN9 produces equally accurate results on par with HNN7 and HNN8. While considering the accuracy only, neural networks from HNN5 to HNN9 are more accurate than others. HNN6 to HNN9 are efficient in terms of precision, recall and F-measure. The overall performance HNN7 and HNN8 is notable and better than other seven networks. The overall performance of HNN6 and HNN9 is also significant when compared with HNN7 and HNN8. The results show that, the number of neurons of the hidden layer should be more than the double of the number of input neurons for this two-pass approach to obtain better classification results.

CONCLUSIONS

A tow-pass artificial neural network approach is proposed and according to that nine hybrid neural networks are built to perform data classification. These nine neural networks are combination of self organizing map and multilayer neural network concepts. In the first pass a set of clusters are created by using Kohonen self organizing map and this cluster results are directed to the multilayer neural network as inputs. The nine networks are of different architecture by means of hidden layer neurons of various multilayer neural networks and named from HNN1 to HNN9. Library users' feedback dataset is used for classification analysis. The performances of the nine neural networks are evaluated by using performance metrics such as accuracy, precision, recall and F-measure. Among the nine, the performance of neural networks from HNN6 to HNN9 is remarkable. The two-pass neural network produces better results if it has at least doubled the number of input neurons in its hidden layer. Increment of neurons in hidden layer influences the performance of two-pass neural networks.

REFERENCES

Hara Brindesi, Maria Monopoli and Sarantos Kapidakis. 2013. "Information seeking and searching habits of Greek physicists and astronomers: a case study of undergraduate students", Procedia - Social and Behavioural Sciences, pp. 785-793.

Keita Tsuji, Erika Kuroo, Sho Sato, Ui Ikeuchi, Atsushi Ikeuchi, Fuyuki Yoshikane and Hiroshi Itsumura. 2012. "Use of Library Loan Records for Book Recommendation", Proceedings of IIAI International Conference on Advanced Applied Informatics, Fukuoka (Japan), pp. 30-35.

Mohd Dasuki Sahak and Mohamad Noorman Masrek. 2013. "Library Usage of Medical Students: A Comparative Analysis of First Year and Third Year Students in University Putra Malaysia", Proceedings of International Conference on Innovation, Management and Technology Research, Penang (Malaysia), pp. 127-132.

Pareek, AK and Madan S. Rana. 2013. "Study of Information Seeking Behaviour and Library Use Pattern or Researchers in the Banasthali University", Journal of Library Philosophy and Practice, pp. 1-9.

Pijitra Jomsri. 2014. "Book Recommendation System for Digital Library Based on User Profiles by Using Association Rule", Proceedings of Fourth International Conference on Innovative Computing Technology, Luton (UK), pp. 130-134.

Ping YU. 2011. "Data Mining in Library Reader Management", Proceedings of International Conference on

Network Computing and Information Security, Guangxi (China), pp. 54-57.

Runhua Wang, Guoquan Liu, Yi Tang, and Yan Li. 2011. "K-means Clustering Algorithm Application in University Libraries", Proceedings of IEEE International Conference on Cognitive Informatics & Cognitive Computing, Alberta (Canada), pp. 419-422.

Runhua Wang, Yi Tang and Lei Li. 2012. "Application of BP Neural Network to Prediction of Library Circulation", Proceedings of 11th IEEE International Conference on Cognitive Informatics & Cognitive Computing, Kyoto (Japan), pp. 420-423.

Sun Yanhua. 2012. "An Assessment Method for College Library Web Site Based on Neural Network", Proceedings of International Conference on Intelligent Systems Design and Engineering Application, pp. 773-775.

Tingting Zhu and Lili Zhang. 2011, "Application of Data Mining in the Analysis of Needs of University Library Users", Proceedings of 6th International Conference on Computer Science and Education (Singapore), pp. 391-394.

Veepu Uppal and Gunjan Chindwani. 2013. "An Empirical Study of Application of Data Mining Techniques in Library System", International Journal of Computer Applications. Vol. 74, pp. 42-46.