Biomedical Research 2016; Special Issue: S87-S91 *ISSN 0970-938X*
www.biomedres.info

Analysis of liver and diabetes datasets by using unsupervised two-phase neural network techniques.

KG Nandha Kumar[1], T Christopher[2*]

[1]Department of Computer Science, Government Arts College, Udumalpet, Tamil Nadu, India

[2]Department of Information Technology, Government Arts College, Coimbatore, Tamil Nadu, India

Abstract

Data classification is a vital task in the field of data mining and analytics. In the recent years, big data has become an emerging field of research and it has wide range of research opportunities. This paper represents three unsupervised and novel neural network techniques: Two-phase neural network (TPNN), stack of TPNN (sTPNN), and ensemble of TPNN (eTPNN) for classification of liver and disbetes data. In this study, Diabetes and Liver data are analyzed by using proposed techniques. Bench mark data sets of liver disorder and diabetes patients records are taken from UCI repository and processed by artificial neural networks towards classification of existence of disease. They are also used for the evaluation of proposed techniques. Performance analysis of three neural classification techniques is done by using metrics such as accuracy, precision, recall and F-measure. sTPNN and eTPNN are found better in overall performance in classifying the disease.

Keywords: Diabetes, Liver disorder, Classification, Data mining, Neural networks.

Accepted on July 30, 2016

Introduction

Data classification is one of the major tasks in the data mining field. Data mining deals with knowledge discovery from large data sets. Cluster analysis and association rule mining are the other parts of data mining. Efficient data mining techniques, methods, algorithms, and tools are playing vital role in bringing unknown facts under the light. Data mining is a field of research which is inevitable for all other fields in this computerized and digital era. Hence various approaches are being used by researchers. Artificial neural network is one of the most famous techniques among research communities which provide better solution for various problems including data classification and clustering. Artificial neural networks are generally referred as neural network or neural net which follows three types of machine learning methods, namely: supervised learning, unsupervised learning and reinforcement learning. Supervised learning techniques are applied for data classification while unsupervised techniques are applied for data clustering normally. But some unsupervised neural network techniques are also found better classifiers when compared with traditional classification algorithms. In this way we have combined the properties of two different unsupervised neural networks self-organizing map and hamming net to construct a neural network classifier to improve the accuracy of classification.

Related work

A novel approach is proposed by Lin et al. [1] to improve the classification performance of a polynomial neural network (PNN) which is also called as a higher order neural network. They have applied real coded genetic algorithm (RCGA) to improve the efficiency of PNN. Ten folds cross validation is performed by the researchers by using Irvine benchmark datasets. Ditzler et al. [2] have developed two deep learning neural network methods for metagenomic classification. A recursive neural network and a deep belief network are implemented and tested with metagenomic data. In the recursive neural network, a tree is used to represent the structure of data. The main tasks are, learning hierarchical structure in a metagenomic sample and classification of phenotypes. It is concluded that traditional neural networks models are more powerful than baseline models on genomic data classification.

Hong-Liang [3] has classified the imbalanced protein data by using ensemble classifier technique EnFTM-SVM. This is an ensemble of fuzzy total margin support vector machine (FTN-SVM). It is a three stage framework. In the first stage, protein feature extraction and representation is made. In the second stage, large numbers of distinct data sets are created. Protein sequences have multiple classes to classify and receiver operating characteristic curve is used to evaluate the classification model. Kumar et al. [4] have introduced a modified technique for the recognition of single stage and multiple power quality disturbances. They have proposed an

algorithm which combines S-Transform based artificial neural network classifier and rule based decision tree. Different types of disturbances to the power quality are classified based on IEEE-1159 standard. To classify power quality events, two-layered feed forward neural network with sigmoid function is used. Scaled conjugate gradient back propagation algorithm is used for network training. After thorough investigation, this proposed algorithm is implemented in real time events and the validity is confirmed.

Wu et al. [5] have proposed a hybrid constructive algorithm for single layer feed forward networks learning (SLFN) which is widely used for classification and regression problems. The SLFN learning has two tasks; determining the network size and training the parameters. The proposed hybrid constructive algorithm can train all the parameters and determine the size of the network simultaneously. In the beginning stage, they have applied the proposed hybrid algorithm which combined Levenberg-Marquardt algorithm and least square method to train the SLFN with fixed network size. Later, they have applied the proposed hybrid constructive algorithm which follows incremental constructive scheme. In this proposed method a new neuron is randomly initialized and added with the network when the training entrapped into local minima problem. Training is continued on previous results with the added new neurons. This hybrid constructive algorithm starts the training with no hidden neurons and increases the hidden neurons one by one every time. The performance and efficiency of this novel algorithm is proved through experiments. Dong et al. [6] have implemented a novel method for vehicle type classification by using semi-supervised convolutional neural network. Sparse Laplacian filter learning is introduced for network training in the convolutional layer and it is an unsupervised method. Beijing Institute of Technology vehicle data set which includes 9850 frontal view images of vehicle is used for the experiments. The neural network classifies the images based on vehicle types such as bus, minivan, truck etc.

Proposed Methods

Three neural network techniques are proposed in this paper. The first one is constructed as basic technique and other two are variants of the first one.

1. Two Phase Neural Network (TPNN)
2. Stack of Two Phase Neural Network (TPNNs)
3. Ensemble of Two Phase Neural Network (TPNNe)

Two phase neural network (TPNN)

A two-phase method is proposed for data classification. In the first phase preprocessed data set is processed by a self-organizing map to find the data clusters and the output vectors are sent to the second phase for effective classification. The working model is represented in Figure 1. Self-organizing map is invented by Kohonen [7] and it is used as a clustering tool and its efficiency is proved by various researchers. It contains two layers namely input layer and output layer. The variable 'n' denotes the number of neurons in input layer and 'm' denotes number of neurons in output layer. The clustering process takes place in the output layer by identifying the winner neuron. The winner neuron is identified by recursively calculating the Euclidean distance to measure the similarity. Since self-organizing map is an unsupervised method, learning takes place at output layer itself and prior training is not required.

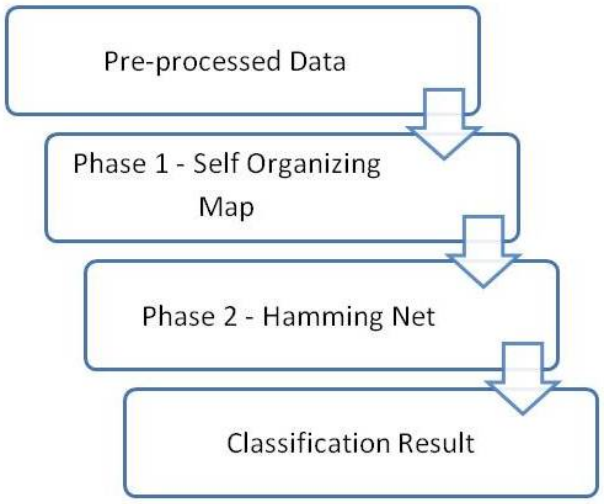

Figure 1. *Two phase neural network for data classification.*

In the second phase the Hamming net will process the cluster results of previous phase. Hamming net is an unsupervised neural network method invented by Lippmann [8] and it is an effective classifier. Here the target is improving the classification efficiency of hamming net through self-organizing map. It finds the exemplar vector by calculating Hamming distance among input vectors and it is determined by number of components in which the vectors differ. In this technique variables b, n, m represent bias value, number of neurons in input layer and number of neurons in output layer respectively. Calculation of Hamming distance is done at first layer. The second layer decides the winner vector which has minimum Hamming distance through a Maxnet as a subnet. Maxnet was developed by Lippmann and it is a subnet which has been constructed by using fixed neural nodes and used to classify the values when large input is given.

First phase (Self Organizing Maps):

Step 1. Initialize the weights and setting topological parameters.

Step 2. Calculate the square of Euclidean distance for each input vector.

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

Step 3. Find the final unit index J, so that D(J) is minimum.

Step 4. Update weights for all j.

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

Step 5. Update the learning rate α by α(t+1)=0.5α (t).

Step 6. Reduce topological parameter and test for stopping condition.

Second phase (Hamming net):

Step 1. Initialize the weights.

For (i=1;i<N;i++); For (j=1;j<M;j++); Wij = ei(j)/2.

Step 2. Initialize the bias. Bj=N/2

Step 3. Calculate the net input. $Yinj = Bj + \sum_{i=1}^{N} x_i w_{ij}$

Step 4. Initialize the activation. Yj(0)=Yinj, j=1 to M

Step 5. Repeat step 3 and step 4 up to finding the exemplar and stop.

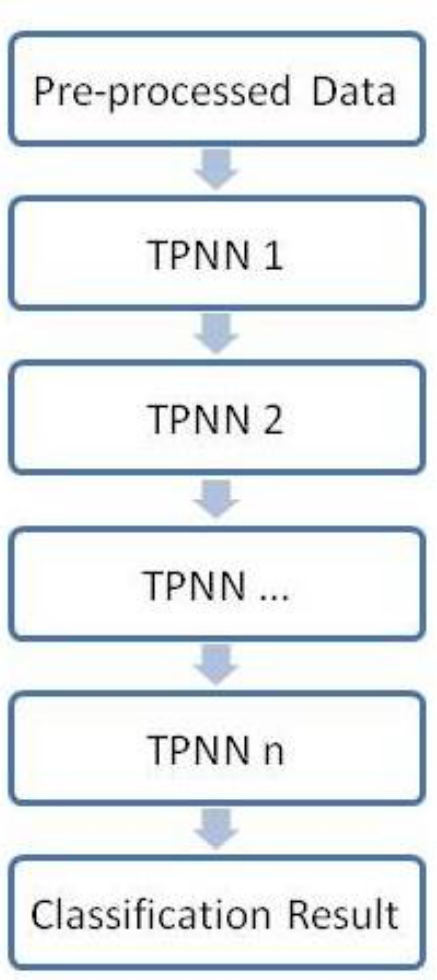

Figure 2. *Stack of two phase neural network for data classification.*

Stack of two phase neural network (sTPNN)

Architecture plays a vital role in any artificial neural networks. Hence to improve the TPNN, two architectures are incorporated with the two phase neural network. The TPNN is enhanced with stack architecture. In stack structure, a neural network model is built through the concept one on another to make a neural network stack as shown in Figure 2. Stack of two phase neural network is a deep network. It performs deep learning through multiple layers.

Ensemble of two phase neural network (eTPNN)

An ensemble of two phase neural networks is constructed to enhance the performance of TPNN. An ensemble is a collection of same or different kind of techniques which is developed to improve the overall performance. In an ensemble, a dataset will be processed by all the members of ensemble and the result of all techniques will be summarized as shown in Figure 3.

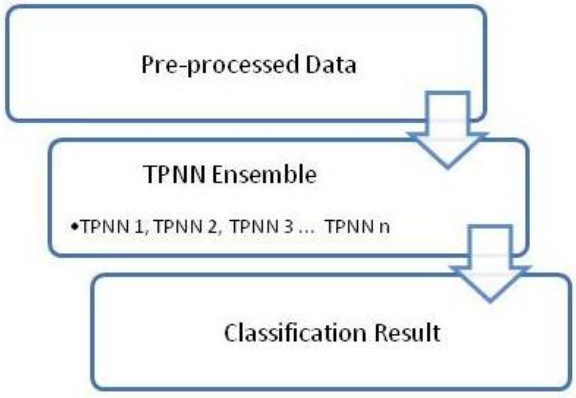

Figure 3. *Ensemble of two phase neural network for data classification.*

Results

All the three neural network techniques have two neural layers basically. They are implemented and tested with three combinations based on number of neurons in input layer and processing layer such as 8-18, 8-20, and 8-22. Based on the previous research and results [9], number of neurons of input layer is fixed as eight and number of neurons of processing layer is fixed as eighteen, twenty and twenty two. The stack of TPNN technique has been implemented and tested by three stacks based on the repetition of TPNN in the particular stack. In the first, the second and the third stacks, TPNN is repeated for three, five and seven times respectively. The same approach is followed in the construction of ensembles too. Three ensembles have been implemented and tested which has three, five and seven TPNNs respectively. All the techniques are tested with Irvine benchmark data sets of University of California and detailed description of datasets are given in Table 1.

Table 1. *Details of datasets.*

Sl. No.	Name of Dataset	No. of Records	No. of Attributes	No. of Classes
1	Australian credit approval	690	14	2
2	BUPA Liver	345	6	2

3	Diabetes	768	8	2

The classification performance of the techniques is measured by the values of accuracy (A), precision (P), recall (R) and F-measure (F) based on classified values as follows.

- A=(TP+TN)/(TP+FP+TN+FN), P=TP/(TP+FP), R=TP/(TP+TN), F= (2*P*R)/(P+R).
- TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative.

Table 2. *Performance evaluation of TPNN.*

Datasets	Combination of TPNN	Accuracy (A)	Precision (P)	Recall (R)	F-measure (F)
Australian credit approval	8-18	82.0%	0.846	0.537	0.657
	8-20	86.0%	0.815	0.512	0.629
	8-22	86.0%	0.920	0.535	0.676
BUPA Liver	8-18	88.2%	0.885	0.511	0.648
	8-20	86.0%	0.880	0.512	0.647
	8-22	86.5%	0.857	0.533	0.658
Diabetes	8-18	87.8%	0.885	0.535	0.667
	8-20	86.0%	0.870	0.465	0.606
	8-22	88.0%	0.846	0.500	0.629

All the three datasets are given as input and processed by all the three neural network techniques. The evaluation details of performance of TPNN, sTPNN, and eTPNN techniques are presented in Tables 2-4 respectively. The TPNN has classified the Australian credit approval data set with 86% accuracy, BUPA Liver data set with 88% accuracy, and Diabetes data set with 88% accuracy. Another notable point is, the increment of number of neurons in processing layer reflects in accuracy level. The increment of processing neurons from 18 to 22 produces very less amount of variation in accuracy hence having such amount of neurons will be feasible for effective learning. The maximum achievement of accuracy, precision, recall, and F-measure of TPPN are 88%, 0.92, 0.537, and 0.676 respectively. The sTPNN has classified the Australian credit approval data set with 88% accuracy, BUPA Liver data set with 90% accuracy, and Diabetes data set with 90% accuracy. From this result it is proved that stack architecture increases the classification accuracy. The maximum achievement of accuracy, precision, recall, and F-measure of sTPPN are 90%, 0.96, 0.548, and 0.696 respectively.

Table 3. *Performance evaluation of sTPNN.*

Datasets	Combination of sTPNN	Accuracy (A)	Precision (P)	Recall (R)	F-measure (F)
Australian credit approval	Stack of 3 TPNNs	84.0%	0.885	0.548	0.676
	Stack of 5 TPNNs	88.0%	0.852	0.523	0.648
	Stack of 7 TPNNs	88.0%	0.960	0.545	0.696
BUPA Liver	Stack of 3 TPNNs	90.2%	0.923	0.522	0.667
	Stack of 5 TPNNs	88.0%	0.920	0.523	0.667
	Stack of 7 TPNNs	88.5%	0.893	0.543	0.676
Diabetes	Stack of 3 TPNNs	89.8%	0.923	0.545	0.686
	Stack of 5 TPNNs	88.0%	0.913	0.477	0.627
	Stack of 7 TPNNs	90.0%	0.885	0.511	0.648

Table 4. *Performance evaluation of eTPNN.*

Datasets	Combination of eTPNN	Accuracy (A)	Precision (P)	Recall (R)	F-measure (F)
Australian credit approval	Ensemble of 3 TPNNs	84.0%	0.846	0.524	0.647
	Ensemble of 5 TPNNs	86.0%	0.815	0.512	0.629
	Ensemble of 7 TPNNs	88.0%	0.920	0.523	0.667
BUPA Liver	Ensemble of 3 TPNNs	88.2%	0.885	0.511	0.648
	Ensemble of 5 TPNNs	88.0%	0.880	0.500	0.638
	Ensemble of 7 TPNNs	86.5%	0.857	0.533	0.658
Diabetes	Ensemble of 3 TPNNs	89.8%	0.885	0.523	0.657
	Ensemble of 5 TPNNs	86.0%	0.870	0.465	0.606
	Ensemble of 7 TPNNs	90.0%	0.846	0.489	0.620

The eTPNN has classified the Australian credit approval data set with 88% accuracy, BUPA Liver data set with 88% accuracy, and Diabetes data set with 90% accuracy. From this result it is proved that ensemble architecture is equally efficient when compared with stack architecture and it increases the classification accuracy. The maximum achievement of accuracy, precision, recall, and F-measure of eTPPN are 90%, 0.92, 0.533, and 0.667 respectively.

Conclusion and Future Scope

Three unsupervised neural network techniques Two-phase neural network (TPNN), stack of TPNN (sTPNN), and ensemble of TPNN (eTPNN) are proposed in this paper for classification of Liver disorder and for classification distebetes problem. BUPA liver and diabetes data sets from UCI repository are used for this study. Australian credit approval data set is also processed and analysed by the proposed

techniques for validation. Performance analysis of three neural network based classification techniques are done by using metrics such as accuracy, precision, recall and F-measure. In terms of accuracy, sTPNN and eTPNN perform well. sTPNN performs well in terms of recall, precision, and F-measure. Among the three techniques, sTPNN and eTPNN are found better in overall performance even though only slight variations found in the performance of three techniques. The sTPNN and eTPNN techniques produce better results on diabetes and liver datasets. They have classified the disorder of liver more accurately than other techniques and also diabetes records also classified properly by the same techniques. It is concluded that architectural changes such as increment & decrement of neurons in a layer, merging of different networks, ensemble learning will improve the disease classification performance of neural network techniques. In this study classification accuracy is improved by using stack and ensemble architectures of artificial neural networks. There are scope for applying other soft computing techniques such as fuzzy logic, genetic algorithms, and hybrid techniques such as neuro-fuzzy, genetic-neuro, genetic-fuzzy for improving the performance of classifiers on medical datas.

References

1. Chin-Teng L, Prasath M, Saxena A. An Improved Polynomial Neural Network Classifier using Real-Coded Genetic Algorithm. IEEE Transact Syst Man Cybernet Syst 2015; 45: 1389-1401.
2. Ditzler G, Polikar R, Rosen G. Multi-layer and Recursive Neural Networks for Metagenomic Classification. IEEE Transact Nanobiosci 2015; 14: 608-616.
3. Hong-Liang D. Imbalanced Protein Data Classification using Ensemble FTM-SVM. IEEE Transact Nanobiosc 2015; 14: 350-359.
4. Kumar R, Singh B, Shahani DT, Chandra A, Al-Haddad K. Recognition of Power-Quality Disturbances using S-Transform-based ANN Classifier and Rule-based Decision Tree. IEEE Transact Industry Appl 2015; 51: 1249-1258.
5. Wu X, Rozycki P, Wilamowski BM. A hybrid constructive algorithm for single-layer feedforward networks learning. IEEE Transact Neural Network Learning Syst 2015; 26: 1659-1668.
6. Dong Z, Wu Y, Pei M, Jia Y. Vehicle Type Classification using a Semisupervised Convolutional Neural Network. IEEE Transact Intell Transport Syst 2015; 16: 2247-2256.
7. Kohonon T. The Self Organizing Map. Proceedings of IEEE 1990; 78: 1464-1478.
8. Lippmann RP. An Introduction to Computing with Neural Nets. IEEE ASSP Magazine 1987.
9. Kumar NKG, Christopher T. A Novel Neural Network Approach to Data Classification. ARPN J Eng Appl Sci 2016; 11: 6018-6021.

***Correspondence to**

Christopher T

Department of Information Technology

Government Arts College

India