Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 4, April 2013, pg.88 – 94

RESEARCH ARTICLE

Supervised Neural Network using Maximum-Margin (MM) Principle

Prakash J. Kulkarni[1], Somnath Kadam[2]
[1]Computer Science & Engineering, Walchand College of Engineering, Sangli, India
[2]Computer Science & Engineering, Walchand College of Engineering, Sangli, India

[1] *pjk_walchand@rediffmail.com;* [2] *somnath.a.kadam@gmail.com*

Abstract— This paper presents a method to build binary classifier using supervised neural network. Support Vector Machine (SVM), which is based on the concept of maximum margin, is also used to build binary classifier, however it is mathematically complex. Neural Network (NN) is a simpler alternative and more suitable for parallel processing. This paper presents Maximum margin algorithm (MMGDX), which stretches out the distance between two classes (margin) to its maximum limit. It uses back-propagation method for error calculation and also uses gradient descent with adaptive learning rate to increase learning rate of network. In MMGDX, area under receiver operating characteristics (ROC) curve (AUC), is applied for stopping criterion. Real time benchmark data sets are used for experiments that enable comparison with other state-of-art classifiers.

Key Terms: - supervised neural networks; classifier; Maximum margin; AUC

I. INTRODUCTION

A. Neural Network

Artificial neural network is an excellent technique to solve a number of problem areas, whereas conventional Von Neumann computer systems have been traditionally slow and inefficient. One of these is pattern recognition. In pattern recognition, there are three main problems associated with it. The very first one is suitable composition for the training data set, second is feature extraction and/or selection and the last one is classifier. This paper mainly focuses on classifier. Classification is an instance of supervised learning, i.e. learning where a training set has input and output pattern.

MMGDX uses multilayer perceptron (MLP) as classifier, and it is universal approximator [7], i.e. MLPs can fit any dataset. Multilayer perceptron using a back-propagation algorithm is the standard algorithm for any supervised learning pattern recognition process.

The back-propagation learning process consists of small iterative steps: one of the examples is applied to the network. Then, network produces output, known as estimated output, based on the current state of its synaptic weights (initially, the weight will be random). This estimated output is compared with the desired output and a mean-squared error signal is calculated. The error is back propagated through the network, and based on that weights of each layer are updated. The whole process is repeated for all cases of an example, then back to the first case again, and so on. The weights are updated for all examples. The whole process is repeated until overall error value is dropped below some threshold. At this stage, network has learned the problem "well enough" and network will never exactly learn the ideal function. The back-propagation algorithm for calculating a gradient has been rediscovered many times.

© 2013, IJCSMC All Rights Reserved

The problem with supervised neural network is an over-fitting problem i.e. an error on training data set is small and it increases when new data is presented. To deal with this problem, MMGDX uses the value of AUC [6] as a stopping criterion for training section. ROC curves are widely used for visualization and comparison of performance of binary classifiers. AUC is a single scalar value for classifier comparison. Statistically, AUC of a classifier is the probability to rank randomly chosen positive instances higher than randomly chosen negative instances.

II. LITERATURE SURVEY

"Beyond feed forward models trained by back propagation: A practical training tool for a more efficient universal approximator": There are many complex neural models, such as Simultaneous recurrent neural networks (SRNs) [8], and MLP both are universal approximator. MLP can approximate an arbitrary nonlinear, continuous, and multi-dimensional function with any desired accuracy.

Because of simplicity of MLP, this model is used for pattern recognition applications. The MLP is used to achieve better performance than (or at least similar) state-of-the-art approaches, such as support vector machines (SVMs) with nonlinear kernel [9], [11], Bayesian neural network [10], or novel algorithms based on kernel Fisher discriminant analysis [12].

"CARVE: A constructive algorithm for real-valued examples": It finds a hyper plane that separates a set of data belonging to one class from the other class of the data. Then, it removes separated data from the training data, and repeats this procedure until only one class of data remain [3] [4].

"Semi-supervised Neural Networks for Efficient Hyper-spectral Image Classification": Training is done using stochastic gradient descent with additional balancing constraints to avoid falling into local minima. The method is useful for supervised and unsupervised methods and can handle millions of unlabelled samples [13].

III. PRESENTED TECHNIQUE

A. Basic framework

Fig. 1 shows basic framework of classifier. It is a non-linear binary classifier for different types of applications. It takes input and output pair (i.e. supervised learning) as input and classify into positive or negative class.

Fig. 1 Basic framework

B. Methodology

In gradient descent, error is equal to target output minus calculated output. However, in presented method (MMGDX), calculation of an error is based on support vector norm, target output and calculated output. For stopping criteria AUC curve information [1] is used.

In MMGDX algorithm, hidden layer and output layer are jointly optimized in single process. The objective function is back-propagated through the output and hidden layers in such a way as to create a hidden output especially oriented towards getting larger margin for output layer separating hyperplane.

Sigmoidal function is used in hidden layer,

$$yh = \varphi(W1.x + b1) \quad (1)$$

$$\hat{y} = W2.yh + b2 \quad (2)$$

Where x is belong to training data, φ sigmoidal function.

For output layer $b2$ is set to 0, because after training section, ROC curve information is taken into account to adjust the classifier threshold.

Distance between two different classes is defined as

$$d = \frac{\hat{y}}{\|w2\|} \quad (3)$$

Error function is defined as

© 2013, IJCSMC All Rights Reserved

$$ei = (yi\sqrt{n} - \frac{\hat{y}}{\|w2\|}) \quad (4)$$

Objective function (J) is defined as follows

$$J = \frac{1}{N}\sum_{i=1}^{N} ei^2 \quad (5)$$

Where N is total number of training examples

AUC is very useful measure of similarity between two classes measuring AUC. In case of data with no ties, all sections of ROC curve are either vertical or horizontal and in case of data with ties, diagonal sections can also be occurred. One drawback of using presented error measures is that AUC calculation is computationally expensive, Since it usually requires full sorting of the measured dataset.

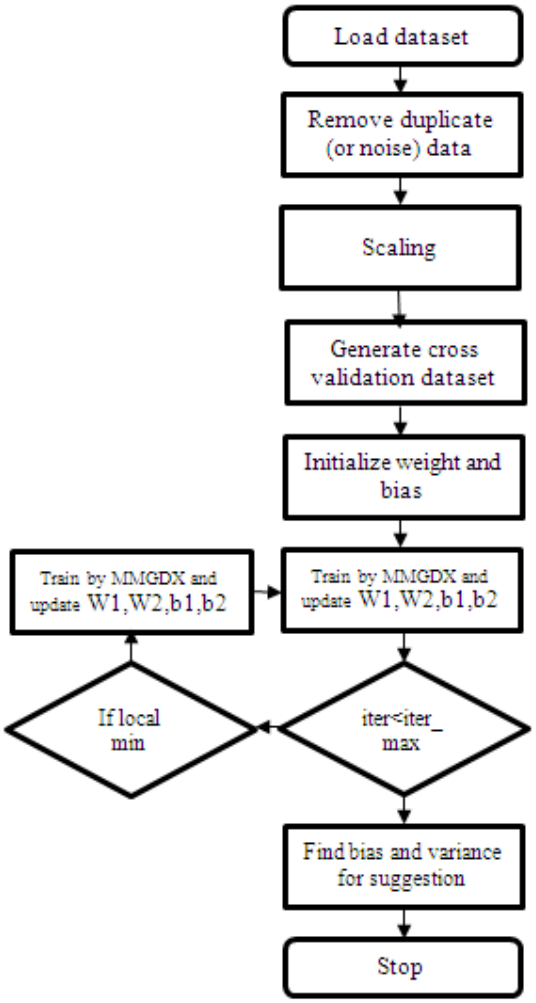

Fig. 2 Flowchart of MMGDX Algorithm

For real-world problems, misclassification costs are unknown and thus, ROC curve and related metrics such as the AUC can be a more meaningful performance measures.

Generalization is the ability to train one data set and to successfully classify independent test sets. Continuous training can increase the training set accuracy, however test set accuracy decreases after a certain point. Over-fitting means error is larger for testing data set. There are different methods for preventing over-fitting like regularization, early stopping criteria, maximal-margin training algorithm and cross validation methods [1] [2] [5].

The bias is equal to how much the averaged overall network output on possible data sets differs from desired function. The variance is equal to how much the network output varies between dataset. The neural network

© 2013, IJCSMC All Rights Reserved

performance can be improved if both the bias and the variance are reduced. However, there is a trade-off between the bias and variance, and it is known as bias/variance dilemma [14].

As per the steps shown in fig. 2, first step is data pre-processing before training of a network. Neural network learns faster and gives better accuracy if the input variables are pre-processed before training of a network. If exactly same pre-processing is done on test dataset, then it avoids unexpected answers from a network. One of the reasons for scaling the data is to equalize the importance of variables.

C. Quality Factor

1) Margin:

- The distance between the separating hyper plane and the training datum nearest to the hyper plane is called the margin.

$$\text{Max. Margin } d = (d+) + (d-) = \frac{2}{||w||} \quad (6)$$

2) Accuracy:

- The accuracy (ACC) is the ration of the total number of predictions, those are correct, to the total number of predictions. It is determined using the equation:
- Table 1 shows confusion matrix entries.
-

Table 1 Confusion Matrix

		Predicate	
Actual		Negative	Positive
	Negative	a	b
	Positive	c	d

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} \quad (7)$$

- "a" is a number of correct predictions of negative class.
- "b" is a number of incorrect predictions of positive class.
- "c" is a number of incorrect predictions of negative class.
- "d" is a number of correct predictions of positive class.

IV. EXPERIMENTAL RESULTS

Algorithm is implemented in MATLAB 7.14 (32-bit) on Windows 8 (32-bit) operating system with Intel i7 processor.

For experiment, new algorithm evaluated two data set [15] shown in table 2.

Thyroid data set: The Thyroid problem task is to decide whether a patient is normal or has a thyroid dysfunction.

Breast data set: Breast-cancer dataset was included in this work due to its difficulty; therefore, this dataset is a suitable option to distinguish the classifier accuracy.

Fig. 3 and fig. 4 are Thyroid data set results for testing and training data sets respectively.

Fig. 5 and fig. 6 are Breast-cancer data set results for testing and training data sets respectively.

Table 2 Data set Information

Data set	**Application**	**Input**	**Output**	**# samples**
Thyroid	Training Set	5	1	140
	Testing Set	5	1	75
Breast-Cancer	Training Set	9	1	200
	Testing Set	9	1	77

© 2013, IJCSMC All Rights Reserved

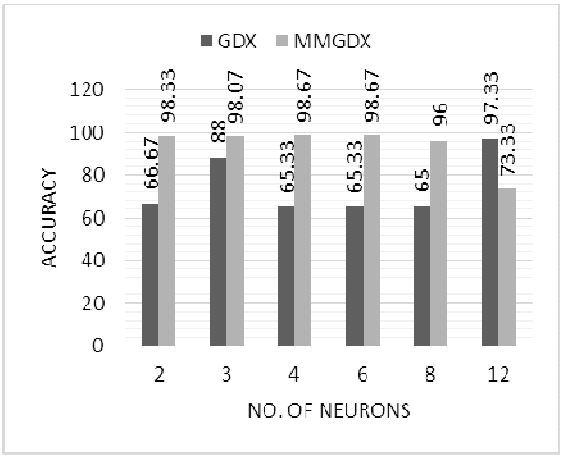

Fig.3 MMGDX applied on Thyroid testing data set

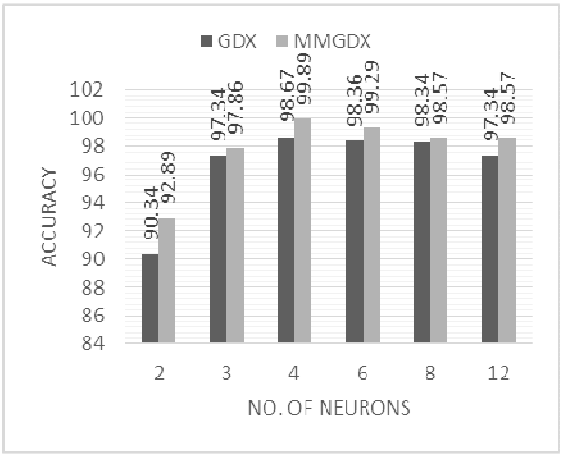

Fig.4 MMGDX applied on Thyroid training data set

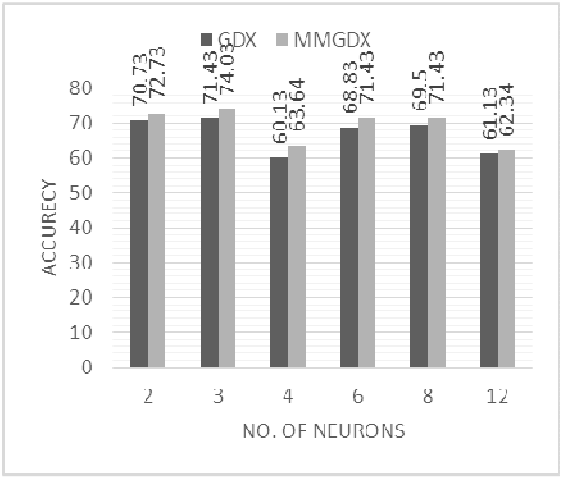

Fig.5 MMGDX applied on Breast-cancer testing data set

© 2013, IJCSMC All Rights Reserved

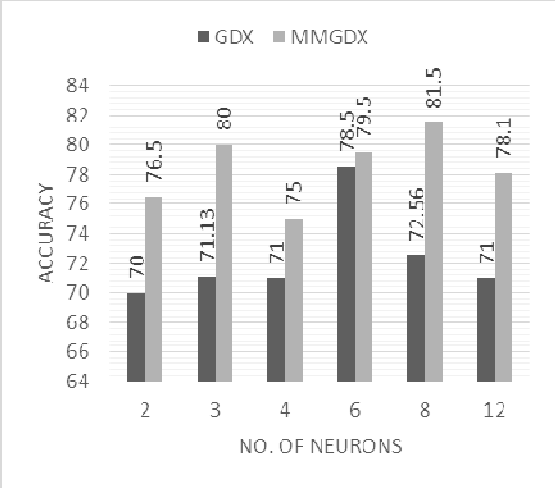

Figure 6 MMGDX applied on Breast-cancer training data set

V. CONCLUSION

The presented MMGDX algorithm provides better accuracy as compared to other state-off-art (GDX) classifier. Results obtained after applying presented MMGDX algorithm on real-world benchmark dataset show better accuracy. Figures (3 to 6) show that MMGDX training is data dependent. So, accuracy of training changes with the data. Accuracy is not increased when neurons are increased above optimal number of neurons.

The presented algorithm also provides solution for over-fitting problem, i.e. error is small on testing and training dataset.

REFERENCES

[1] Oswaldo Ludwig and Urbano Nunes," Novel Maximum-Margin Training Algorithms for Supervised Neural Networks," in IEEE Trans. Neural Netw., VOL. 21, NO. 6, JUNE 2010, pp-972-984.
[2] Shigeo Abe "Support Vector Machines for Pattern Classification" Springer.
[3] S. Young and T. Downs, "CARVE: A constructive algorithm for real-valued examples," IEEE Trans. Neural Netw., vol. 9, no. 6, pp 1180–1190, Nov. 1998.
[4] T. Nishikawa and S. Abe, "Maximizing margins of multilayer neural networks," in Proc. 9th Int. Conf. Neural Inf. Process., Singapore, No 2002, vol. 1, pp. 322–326.
[5] A. K. D. Jayadeva and S. Chandra, "Binary classification by SVM based tree type neural network," in Proc. Int. Conf. Neural Netw, May 2002, vol. 3, pp. 2773–2778.
[6] U. Franke and S. Heinrich, "Fast obstacle detection for urban traffic situations," IEEE Trans. Intell. Transp. Syst., vol. 3, no. 3, pp. 173–181, Sep. 2002.
[7] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," Neural Netw., vol. 3, no. 5, pp. 551–560, 1990.
[8] R. Ilin, R. Kozma, and P. J. Werbos, "Beyond feedforward models trained by backpropagation: A practical training tool for a more efficient universal approximator," IEEE Trans. Neural Netw., vol. 19, no.6, pp. 929–937, Jun. 2008.
[9] C. Liu, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 5, pp. 572–581, May 2004.
[10] R. M. Neal and J. Zhang, "High dimensional classification with Bayesian neural networks and Dirichlet diffusion trees," in Studies in Fuzziness and Soft Computing. Berlin, Germany: Springer-Verlag, 2006, vol. 207, pp. 265–296.
[11] A. Ruiz and P. E. Lopez-de-Teruel, "Nonlinear kernel-based statistical pattern analysis," IEEE Trans. Neural Netw., vol. 12, no. 1, pp. 16–32, Jan. 2001.
[12] J. Yang, A. F. Frangi, J. U. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition," IEEE Trans. pattern Anal. Mach. Intell., vol. 27, no. 2, pp. 230–244, Feb. 2005.
[13] F. Ratle, G. Camps-Valls, and J. Weston, "Semi-supervised neural networks for efficient hyper-spectral image

© 2013, IJCSMC All Rights Reserved

classification," IEEE Trans. Geosci. Remote Sens., vol. 48, no. 5, pp. 2271–2282, May 2010.
[14] Struart Geman , "Neural Network and the bias/variance dilemma", Neural Computation 4,1-58, Massachusetts Institute of Technology 1992.
[15] http://www.nipsfsc.ecs.soton.ac.uk/datasets/

© 2013, IJCSMC All Rights Reserved