ISTANBUL UNIVERSITY –
JOURNAL OF ELECTRICAL & ELECTRONICS ENGINEERING

YEAR : 2004 (1137-1140)
VOLUME : 4
NUMBER : 2

A DATA SELECTION METHOD FOR PROBABILISTIC NEURAL NETWORKS

Bülent BOLAT [1] **Tülay YILDIRIM [2]**

[1,2] Elektronik ve Haberleşme Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
Beşiktaş, İstanbul 34349 TURKEY

[1]E-mail: bbolat@yildiz.edu.tr [2]E-mail: tulay@yildiz.edu.tr

ABSTRACT

In this paper, two performances increasing methods for datasets which have a nonuniform class distribution are presented. The methods are applied to probabilistic neural networks (PNN). Selection of a good training data is the most important issue. Therefore, a new data selection procedure including data exchange and data replication is proposed. After reaching the best accuracy by using the data exchange method, a data replication method is applied to the classes which have relatively less numbers of instances. The methods are applied to the Glass, Escheria Coli (E. coli) and Contact Lenses datasets, which have nonuniform class distributions and better accuracies than the reference works were achieved by PNN using these methods.

Keywords: : *Data selection, performance increasing, PNN.*

1. INTRODUCTION

Data selection is the most important problem for obtaining a good training set to increase the performance of a neural network. A new data selection method based on data exchange and data replication is proposed in this paper for Probabilistic Neural Networks (PNN). The method is applied to Glass, E. coli and Contact Lenses [1] benchmarks and results are given to demonstrate the performance.

Another important problem occurs when a database has at least one class which has relatively low number of instances. Consider a dataset which has 2 classes, 98% of the database belongs to class 1. Accuracy of a neural network which may classify all of the data as class 1 is 98%, however, the class 2 may never be recognised by this network. The methods described below give a good solution for this kind of problem.

2. TRAINING SET SELECTION METHOD

The classifying process begins with finding the optimal spread value for PNN. The optimum spreads are found by a trial-and-error process for all datasets.

Second step of the classifying process is to find a good training set which can give a good accuracy both in training and testing. In this work, an instance exchange method is proposed to choose

Received Date: 19.03.2002
Accepted Date: 15.06.2004

the best training group. The process starts with a random selected training set. After first training process, the test data is applied to the network. A randomly selected true classified instance in the training set (I_1) is thrown into the test set, a wrong classified instance in the test set (I_2) is put into the training set and the network re-trained. If I_2 is false classified, it is marked as a "bad case" and I_1 and I_2 are put into the original locations. If I_2 is true classified and the test accuracy is reduced or not changed, I_1 is put into the original location and another true classified training instance, say I_3, is put into the test set and the process is repeated. If the accuracy is improved, the exchange process is applied to another training and test pairs. The process is repeated until reaching the maximum training and test accuracy.

3. DATA REPLICATION METHOD

After finding the best training group, classes which have relatively low number of instances are considered. All datasets discussed above have nonuniform class distributions. As an example, class 2 of the lenses dataset has only 5 instances but class 3 has 15. Since the 33% of entire dataset is used as test set, it is hard to obtain a good accuracy if the network is trained by only three training data. For this reason, a data replication process is applied to raise the performance of the network. Replication is applied to the first class which has low number of instances and repeated until reaching the best total accuarcy. Then, the process is applied to other classes which have less instances.

4. DESCRIPTION OF THE DATASETS

4. 1. Glass Data Benchmark

Glass dataset [1] was created in the Central Research Establishment, Home Office Forensic Science Service Reading, Berkshire. The dataset has 214 instances separated into six classes. Each instance in the dataset is identified by an id number, nine chemical measurements (where Ri: refractive index, Na: sodium, Mg: magnesium, Al: aluminum, Si: silicon, K: potassium, Ca: calcium, Ba: barium and Fe: iron. All measurements are weight percent in corresponding oxide except refractive index) and a class number between 1 and 7. Class number 4 is reserved and never used in this dataset. Distribution of the dataset is given in the Table 1. Table 2 shows the recent works based on the glass data.

4. 2. Escheria Coli Data Benchmark

Escheria coli is a bacterium; some kinds of E.coli have powerful toxic for human and animal health. The dataset [1] in this paper were taken from Nakai and maintaned by Horton in 1996. E.coli dataset has 336 instances divided into 8 classes. Each instance is identified by a sequence name, eight attributes (where mcg: McGeoh's method for signal sequence recognition, gvh: Von Heijne's method for signal sequence recognition, gvh: Von Heijne's signal peptisade II consensus sequence score, chg: presence of charge on N-terminus of predicted lipoproteins, aac: score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins, alm1: score of the ALOM membrane spanning region prediction program and alm2: score of the ALOM program after excluding putative cleavable signal regions from the sequence.) and a class name. Table 3 shows the class distribution of the dataset, and Table 4 shows the recent works based on this dataset.

4.3. Lenses Data Benchmark

Lenses dataset [1] is a complete database, all possible combinations of attribute-value pairs are represented. The dataset has only 24 instances which are represented by an id number, four attributes which are age (1: young, 2: pre-presbyopic, 3: presbyopic), spectacle prescription (1: myope, 2: hypermetrope), astigmatic (1: no, 2: yes), tear production rate (1: reduced, 2: normal) and a class number. Class 1 has 4, class 2 has 5 instances. The remaining 15 instances are belong to class 3.

5. RESULTS

The simulations were realised by using MATLAB 6.5 Neural Network Toolbox. Training and test set distributions of the databases and replication rates are shown in the Table 5. The numbers between parantheses are replication rates. The accuracies before and after data replication are given in the Table 6. Testing accuracy of the E.coli dataset was not raised by applying data replication, but training accuracy of the class 6 was raised from 92.31% to 100%.

By using data replication on the Lenses dataset, training accuracy of the class 1 was improved from 66.7% to 100%, testing accuracy of the class 2 was increased from 0% to 100% and testing accuracy of the class 3 was reduced from 100% to 80%, but the total testing accuracy of this dataset was raised from 75% to 87.5%. By applying only the data exchange method, the network was fully classified the relatively little classes of the Glass dataset, therefore the data replication was not applied to this dataset. 70% of this dataset was used as training data.

6. CONCLUSION

A new data selection procedure including two performance increasing methods for datasets which have a nonuniform class distribution was proposed in this paper. Data exchange and data replication were used to increase the performance of PNN and these two methods were together or separately applied to some benchmark datasets to make comparison. Much better results than previous works for Glass and E. coli datasets were obtained using these methods.

REFERENCES

[1] http://www.ics.uci.edu/~mlearn/MLRepository.html , 2003.

[2] G. Agre and I. Koprinska, "Case-based Refinement of Knowledge Based Neural Networks", *Proc. of the International Conference on Intelligent Systems: A Semiotic Perspective*, Gaithersberg, MD, USA, pp. 221-226, 1996.

[3] D. Ventura and T. R. Martinez, "An Empirical Comparison of Discretization Methods", *Proc. Tenth International Symposium on Computer and Information Sciences*, pp. 443-450, 1995.

[4] A. Holst, *The Use of a Bayesian Neural Network Model for Classification Tasks*, Dissertation Thesis at Stockholm University, 1997.

[5] H. Ruda and M. Snorasson, "Adaptive Preprocessing for On-line Learning With Adaptive Resonance Theory (ART) Networks", *IEEE Workshop on Neural Networks for Signal Processing*, Cambridge, Massachusetts, USA, 1995.

[6] M. Avci and T. Yildirim, "Classification of Eschericia Coli Bacteria by Artificial Neural Networks", *IEEE International Symposium on Intelligent Systems*, Varna, Bulgaria, Vol: 3, pp. 16-20, 2002.

[7] P. Horton and K. Nakai, "A Probabilistic Classification System for Predicting the Cellular Localization Sites of Proteins", *Intelligent System in Molecular Biology*, Vol: 4, pp. 109-115, 1996.

[8] P. Horton and K. Nakai, "Better Prediction of Protein Cellular Localization Sites With the k Nearest Neighbors Classifier", Intelligent Systems for Molecular Biology*, Vol: 5, pp. 147-152, 1997.*

TABLE I. CLASS DISTRIBUTION OF GLASS DATABASE

Class 1	Class 2	Class 3	Class 5	Class 6	Class 7	Total
70	76	17	13	9	29	214

TABLE II. RECENT WORKS BASED ON GLASS DATABASE

Author	Network Type	Accuracy (%)
Agre and Koprinska [2]	Correction NN by case	68.3
	1-NN	78.8
Ventura and Martinez [3]	C4.5	68
Holst [4]	**Bayesian EM**	87.7
Ruda and Snorasson [5]	ART	Less then 65

TABLE III. CLASS DISTRIBUTION OF ECOLI DATABASE

Class 1 (cp)	Class 2 (im)	Class 3 (imS)	Class 4 (imL)	Class 5 (imU)	Class 6 (om)	Class 7 (omL)	Class 8 (pp)	Total
143	77	2	2	35	20	5	52	336

TABLE IV. RECENT WORKS BASED ON ECOLI DATABASE

Author	Network Type	Accuracy (%)
Avci and Yildirim [6]	PNN	82.97
Horton and Nakai [7]	k-nearest neighborhood	86
Horton and Nakai [8]	Ad hoc probability	81

TABLE V. TRAINING DATA DISTRIBUTION OF DATASETS

E. coli Dataset							
Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
95	51	1	1	23	13 (2)	3	35

Lenses Dataset		
Class 1	Class 2	Class 3
3 (2)	3 (2)	10

TABLE VI. TRAINING AND TEST ACCURACIES

Database	Exchange only (%)		Exchange and Replication (%)	
	Training	Test	Training	Test
Lens	93,33	75	%100	%87,50
E.coli	94,60	90,35	%95,50	%90,35
Glass	98,67	95,31	-	-