An Experimental Decision of Samples for RBF Neural Networks

HYONTAI SUG
Division of Computer and Information Engineering
Dongseo University
Busan, 617-716
REPUBLIC OF KOREA
hyontai@yahoo.com http://kowon.dongseo.ac.kr/~sht

Abstract: - It is not known to decide a proper sample size for data mining tasks, so the task of deciding proper sample sizes for RBF neural networks that are one of the important data mining algorithms tend to be arbitrary. In RBF networks as the size of samples grows, the improvement in error rate becomes better slowly. But we cannot use larger and larger samples, because there are some fluctuations in accuracy as the sample size grows. This paper suggests an objective approach in determining proper samples to find good RBF networks with respect to accuracy. Experiments with two relatively large data sets showed very promising results.

Key-Words: - neural networks, RBF networks, sampling technique

1 Introduction

For the tasks of prediction in data mining neural networks have been very widely used, so neural networks with the smallest error rates for a given data set has been a major concern for their success. But even though neural networks are one of the most successful data mining or machine learning methodologies, there are some points of improvement with respect to accuracy due to the fact that they are built based on greedy algorithms and the knowledge of experts.

In order to train connection weights of neural networks backpropagation algorithms are used, and the backpropagation algorithms rely on some greedy search algorithms like gradient decent search algorithm. So, there is some possibility of considering local optima as global optima.

Even though there are many algorithms to determine the structure of the neural networks, basically the structure of the networks is usually decided by the knowledge of human experts with some experiments. As a result, built neural networks may not represent the best knowledge models that are best for the target domain of the application.

Moreover, because most target databases for data mining are very large, we need sampling process to the target databases. But we know that the task of determining proper sample sizes is arbitrary and the found knowledge based on the random samples is prone to sampling errors or sampling bias.

According to statistics a proper sample size for a feature is 30 or so [1]. For example, to determine the average weight of people, we need to do random sampling for 30 people or so. But, in general, the target databases of data mining contain a lot of features, so if we do sampling like this, the sample size could become enormous. Moreover, according to experiments using RBF networks the accuracy of the trained RBF networks does not increase monotonically as the sample size grows. So, adapting larger and larger sized samples might be of no use to find better RBF networks. Therefore, we need an alternative strategy for sampling.

In section 2, we provide the related work to our research, and in sections 3 we present our method. Experiments were run to see the effect of the method in section 4. Finally section 5 provides some conclusions.

2 Related Work

Neural networks are widely used for machine learning or data mining tasks since the first neural network algorithm, the perceptron [2]. Because of the limited predictability of the perceptron, multilayer perceptrons have been invented. In multilayer perceptrons there are two kinds of networks based on how the networks are interconnected – feed-forward neural networks and recurrent neural networks [3]. Radial basis function (RBF) networks are one of the most popular feed-forward networks [4]. Even though RBF networks have three layers including the input layer, they differ from a multilayer perceptron, because in RBF networks the hidden units perform

some computation. A good point of RBF networks is that they can be trained in relatively rapid speed.

There is also research on sample size [5, 6] as well as the property of samples [7] and sampling method [8, 9]. In [5] the effect of sample size is discussed for parameter estimates in a family of functions for classifiers. In [6] the small sized samples are preferred for feature selection and error estimation for several classifiers. In [7] the authors showed that class imbalance in training data has effects in neural network development especially for medical domain. In [8] Jensen and Oates investigated three sampling schemes, arithmetic, geometric, and dynamic sampling for decision tree algorithms. In arithmetic sampling and geometric sampling, the sample size grows in arithmetic and geometric manner respectively. Dynamic sampling method determines the sample size based on dynamic programming. They found that the accuracy of predictors increases as the sample size increases and the curve of accuracy is logarithmic, so they used the rate of increase in accuracy as stopping criteria for sampling. In [9] several resampling techniques like cross-validation, the leave-one-out, etc. are tested to see the effect of the sampling techniques in the performance of neural networks, and discovered that the resampling techniques has very different accuracy depending on feature space and sample size.

3 The Method

Because we have only limited number of data and the data should be divided into two parts, training and testing, it is not easy to determine an appropriate size of samples that is the best for the target data set. So we resort to repeated sampling technique with various sizes to find the best one. We do the sampling until the sample size is less than the half of the target data set, because we assume that we have some large target data set and we want to have enough test data also. The following is a brief description of the procedure of the method.

INPUT: a data set for data mining,
k: the number of random sampling for each sample size,
s: initial sample size.
OUTPUT: A, V, I, D.
j := 1;
Do while s < | target data set | / 2
Do for i = 1 **to** k /* generate k RBF networks for each loop*/
Do random sampling of size s;
Train and test a RBF network;
a_{ij} := Accuracy of the RBF network;
$A_j := A_j \cup \{a_{ij}\}$;
End for;
$A := A \cup A_j$;
v := the average accuracy in A_j;
$V := V \cup \{v\}$; /* V: average accuracy values */
i := (the average accuracy of the RBF networks of previous step) – (the average accuracy of the RBF networks); /* average improvement rate */
$I := I \cup \{i\}$; /* I: set of i values */
d := (maximum of the accuracy values among the trained RBF networks) - (minimum of the accuracy values among the trained RBF networks);
/* d stands for the fluctuation of accuracy values in the trained RBF networks */
$D := D \cup \{d\}$; /* D: set of d values */
If s >= mid_limit **Then**
s := s + sample_size_increment; j++;
Else
$s := s \times 2$; j++;continue; /* while loop */
End if
End while;

In the algorithm we double the sample size until the size reaches some point, mid_limit, then we increment the sample size with some fixed value, sample_size_increment, because doubling the sample size can exhaust the data soon.

Even though we do random sampling, because we may have some sampling bias and sampling errors as well as due to the property of neural networks, the trained neural networks have a variety in accuracy. So, in order to get rid of the effect of variety in accuracy we average the accuracies of the trained neural networks for each sample size, and this average accuracy with improvement value and fluctuation value in accuracy is used to determine a proper sample size. By selecting a sample size that generates good RBF networks in average with satisfactory accuracies, we can have better RBF networks in predictability in future cases.

4 Experimentation

Experiments were run using data sets in UCI machine learning repository [10] called 'census-income' and

'adult' to see the effect of the method. Adult data set is a refined version of census-income data set. The number of instances in census-income data set for training is 199,523 in size of 99MB data file. The number of instances in adult data set is 48,842. The data sets were selected, because they are relatively very large and contain lots of values. The total number of attributes is 42 and 14, and among them eight and six attributes are continuous attributes for census-income and adult respectively. The values in continuous attributes of census-income data set are converted to nominal values with entropy-based discretization method, because the method showed the best result according to the experiments in [11].

We used RBF network using logistic regression applied to K-means clustering to train for various sample sizes. The following Table 1 and 2 show average accuracy depending on various sample sizes for census-income and adult data set respectively. For each sample size seven random samples have been selected and seven neural networks have been generated for the experiment.

The initial sample size for training is 2,000 and 200 for census-income and adult respectively, and the size of samples is doubled as the while loop runs. For census-income and adult the given mid_limit value for sample size are 16,000 and 6,400 respectively, and the sample_size_increment of 8,000 and 3,200 for census-income and adult respectively. The rest of the data set after sampling is used for testing.

In the table, the fourth column, improvement(%), means the percentage of improvement in accuracy compared to the neural networks of previous sample size, and the last column represents the difference of maximum and minimum values of accuracy among the RBF networks in the given sample size.

Table 1. RBF networks for 'census-income' data set with various sample sizes

Samp. Size	Average accuracy(%)	Improve -ment(%)	Diff. of max & min accuracy(%)
2,000	94.12973	NA	0.6957
4,000	94.10299	-0.02674	0.5974
8,000	93.97587	-0.12712	0.7122
16,000	93.96534	0.01053	0.674
24,000	*94.21419*	*0.24885*	*1.12391*
32,000	94.11256	-0.10163	0.6196
40,000	94.05337	-0.05919	0.6833
48,000	*94.30241*	*0.24904*	*1.1826*
56,000	94.10687	-0.19554	0.9964
64,000	94.12129	0.01442	0.9637

If we look at table 1, sample size 48,000 has the best accuracy, and the secondly best is sample size 24,000. The best accuracy in sample size 24,000 is 95.0177% and the best accuracy in sample size 48,000 is 94.9367% so that we may choose one of them as our neural network. Note that as the sample size increases, accuracy does not increase monotonically.

Table 2. RBF networks for 'adult' data set with various sample sizes

Samp. size	Average accuracy(%)	Improve -ment(%)	Diff. of max & min accuracy(%)
200	82.15153	NA	2.4239
400	83.3527	1.20117	1.6907
800	82.86174	-0.49096	0.9783
1,600	83.13183	0.27009	1.5071
3,200	*83.64977*	*0.51794*	*1.1419*
6,400	83.38611	-0.26366	2.0288
9,600	*83.57734*	*0.19123*	*0.6345*
12,800	83.45717	-0.12017	0.6165

If we look at table 2, sample size 9,600 has the best accuracy, and the secondly best is sample size 3,200. The best accuracy in sample size 3,200 is 84.1506% and the best accuracy in sample size 9,600 is 83.8846% so that we may choose one of them as our neural network. Note that as the sample size increases, accuracy does not increase monotonically also. Note also that bigger sample sizes have less fluction in difference of maxmum and minimum accuracy values for adult data set.

5 Conclusions

Neural networks are widely accepted for data mining or machine learning tasks and it is known that neural networks are one of the most successful data mining tools for prediction. But, neural networks may not always be the best predictors due to the fact that they are trained based on some greedy algorithms with limited data sets and the knowledge of experts. So, some improvements may be possible.

Because the target data sets in data mining tasks contain a lot of data, random sampling has been considered a standard method to cope with large data sets that are very common in data mining task. But, simple random sampling might not generate perfect samples that are good for the used data mining algorithms. Moreover, the task of determining a

proper sample size is arbitrary so that the reliability of the trained data mining models might not be good models to be trusted.

We propose a repeated sampling method with various sample sizes to decide the best random samples for RBF networks that are one of the good neural network algorithms for data mining. Experiments with real world data sets showed very promising results.

References:

[1] W.G. Cochran, *Sampling Techniques*, 2nd ed., Wiley, 1997.

[2] M.L. Minsky, S.A. Papert, *Perceptrons – extended edition: an introduction to computational geometry*, MIT press,1987.

[3] P. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Addison Wesley, 2006.

[4] C.M. Bishop, *Neural networks for pattern recognition*, Oxford University press, 1995.

[5] K. Fukunaga, R.R. Hayes, Effects of Sample Size in Classifier Design, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11, issue 8, 1989, pp. 873 - 885.

[6] S.J. Raudys, A.K. Jain, Small Sample Size Effects in Statistical Pattern recognition: Recommendations for Practitioners, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, No. 3, 1991, pp. 252—264.

[7] M.A. Mazuro, P.A. Habas, J.M. Zurada, J.Y. Lo, J.A. Baker, G.D. Tourassi, Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, Neural Networks, Vol. 21, Issues 2-3, 2008, pp. 427 – 436.

[8] T. Oatesm, D. Jensen, "Efficient progressive sampling", Proceedings of the Fifth International Conference on Knowledge Discovery and data Mining, 1999, pp. 23-32.

[9] S. Berkman, H. Chan, L. Hadjiiski, Classifier performance estimation under the constraint of a finite sample size: Resampling scheme applied to neural network classifiers, Neural Networks, Vol. 21, Issues 2-3, 2008, pp. 476 – 483.

[10] D. Newman, *UCI KDD Archive* [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science, 2005.

[11] Liu, H., Hussain, F., Tan, C.L., Dash, M., "Discretization: An Enabling Technique", *Data Mining and Knowledge Discovery*, Vol. 6, 393-423, 2002.