Property of Artificial Neural Networks of Classification with respect to Training Set Size

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: - Multilayer perceptrons and radial basis function networks are used most often in classification tasks, even though the two neural networks have different performance in classification tasks depending on the available training data sets. This paper shows the accuracy change in classification of the two neural networks when training data set size changes. Experiments were run with four data sets and found that multilayer perceptrons show relatively better accuracies as the training data set size grows, while radial basis function networks do not improve much compared to the other neural networks, as a result the accuracy of multilayer perceptron became better for two data sets, as the training data set size grew.

Key-Words: - artificial neural networks, the size of training data, prediction accuracy

1 Introduction
Artificial neural networks have been very successful for the task of data mining or machine learning [1, 2] and recent development in radial basis function (RBF) networks draws many researchers’ attention, because many researchers have reported their good performance in various application areas.

Because classification accuracy is an important factor for data mining, developing neural networks with the smallest error rates for a given data set have been a major concern for their success. Even though multilayer perceptrons (MLPs) and RBF networks are two major neural networks that have been applied successfully for classification tasks, their training mechanisms are very different. MLPs use backpropagation algorithms to train the connection weights so that it takes long time to train, because the backpropagation algorithms rely on some greedy search algorithms like gradient decent. Even though the gradient descent works well in most cases, there is still some possibility of considering local optima as global optima [3]. In structural aspect, RBF networks have three layers including the input layer, hidden layer, and output layer, they differ from a MLP, because in RBF networks have radial basis functions that are trained by k-means clustering in general in the hidden layer [4, 5].

Because the training algorithms of the neural network are based on their own kinds of induction, it is also true that the performance of neural networks is dependent upon the available data sets. So, in this paper we want to see any relationship between the accuracy of neural networks, especially for MLPs and RBF networks, and the data set size.

In section 2, we provide the related work to our research, and in sections 3 we present our method of experiment, and experiments were run to see the property of the two neural networks in section 4. Finally section 5 provides some conclusions.

2 Related work
Artificial neural networks are divided into two kinds of networks based on how the networks are interconnected – feed-forward neural networks and recurrent neural networks [6]. RBF networks are one of the most popular feed-forward networks. Papers like [7, 8, 9, 10, 11] show better performance of RBF networks than other neural network algorithms, while some papers prefer MLPs [12]. In RBF networks, local optima problem may occur due to the feed-forward nature of the network and structure of the hidden layer. In order to overcome this problem many evolutionary search algorithms were suggested [13, 14, 15]. Evolutionary search algorithms try to find global optimal solutions so that it is possible to find better RBF networks. But the algorithms require more extensive computing time as well as more elaborate techniques related to the evolutionray computation like the representation technique of network structures and weights.

Because the behavior of trained knowledge models also dependent on the training data set, there is research on sample size as well as the property of samples and
sampling scheme. In paper [16] the authors discussed the effect of sample size for parameter estimates in a family of functions for classifiers. In paper [17] the authors prefer small sized samples for feature selection and error estimation for several classifiers of pattern recognition. In [18] the authors showed that class imbalance in training data has effects in neural network development especially for medical domain. In paper [19] the authors 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 their progressive sampling method.

3 The method of experiment
We want to check the prediction accuracy of RBF networks and MLPs with changing training data set size to find out the effect of training data set size for the performance of the two neural networks. We increment the size of training data set size in arithmetic or geometric progress depending on the available data set size.

The following is a brief description of the procedure of the experiment.

```
INPUT: a data set
s: initial sample size.

OUTPUT: R, M
/* R: the set of accuracy of RBF networks,
M: the set of accuracy of MLPs */
i := 1;
Do While s = |target data set| / 2
    Repeat m times
        Do random sampling of size s;
        Train and test RBF network and MLP;
        \(r_i := \) the accuracy of the RBF network;
        \(m_i := \) the accuracy of the MLP;
        \(R_i := R_{i-1} \cup \{r_i\}\);
        \(M_i := M_{i-1} \cup \{m_i\}\);
        Increment s in arithmetically or geometrically;
    End repeat;
End while;
```

In the above algorithm we increment the sample size until the sample size reaches to about half of the data set size. The used RBF networks and MLPs should have appropriate network structures and appropriate number of iterations for back propagation based on the characteristics of the given data sets.

4 Experimentation
Experiments were run using four data sets in UCI machine learning repository [20] called ‘ozone(eight hour)’, ‘yeast’, ‘census’, and ‘statlog’ to see the effect of the method. In previous work [21] ‘adult’ and ‘forest cober types’ data sets are used [21].

RBF network of k-means clustering and MLPs are used for the experiment [22]. Table 1 to 4 show the result of training for the two neural network algorithms.

For ‘ozone’ and ‘yeast’ data sets 2 was given for parameter of clusters for RBF networks, and the number of hidden layers for MLPs is half of the number of attributes plus the number of classes, and training time is 500. The number of instances in ‘ozone’ and ‘yeast’ data set is 2,536, and 1,484 respectively. The initial sample size for training is 200, and the rest of the data set after sampling is used for testing, so we have bigger test set data when sample size is small. In the experiment four random samples for each sample size are used. The tables contain average accuracy values.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Accuracy of RBFN(%)</th>
<th>Accuracy of MLP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>93.4983</td>
<td>92.1165</td>
</tr>
<tr>
<td>400</td>
<td>93.7551</td>
<td>92.58133</td>
</tr>
<tr>
<td>800</td>
<td>93.6563</td>
<td>92.67365</td>
</tr>
<tr>
<td>1,200</td>
<td>93.497</td>
<td>93.12278</td>
</tr>
<tr>
<td>1,600</td>
<td>93.60635</td>
<td>93.73985</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Accuracy of RBFN(%)</th>
<th>Accuracy of MLP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>46.80688</td>
<td>53.0179</td>
</tr>
<tr>
<td>400</td>
<td>52.89413</td>
<td>54.8774</td>
</tr>
<tr>
<td>600</td>
<td>55.20363</td>
<td>58.20135</td>
</tr>
<tr>
<td>800</td>
<td>56.4488</td>
<td>58.1289</td>
</tr>
</tbody>
</table>

If we look at table 1 and 2, we can notice that RBFNs are better for ‘ozone’ data set, while MLPs are better for ‘yeast’ data. So, we can conclude that the two target data sets have different favor for each neural network.
total number of attributes including class attribute is 42. Among them eight attributes are continuous attributes. The census income data set has very big data records, and the size of the data set is very large, so, 42 were chosen as the parameter for clusters for K-means clustering. In order to train MLPs the given number of hidden layers is ten. Table 3 shows the result of training for the two neural network algorithms. The initial sample size for training is 2,500, and the size of samples is doubled as the while loop runs, and we stop sampling when the sample size reaches to about 20% of the data set size due to long computing time.

Table 3. RBFNs and MLPs for ‘census income’ data set with different sizes of training data set

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Accuracy of RBFN(%)</th>
<th>Accuracy of MLP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,500</td>
<td>93.91588</td>
<td>94.02115</td>
</tr>
<tr>
<td>5,000</td>
<td>94.3767</td>
<td>94.19763</td>
</tr>
<tr>
<td>10,000</td>
<td>94.33915</td>
<td>94.0922</td>
</tr>
<tr>
<td>20,000</td>
<td>94.35875</td>
<td>94.62895</td>
</tr>
<tr>
<td>40,000</td>
<td>94.48245</td>
<td>94.6336</td>
</tr>
<tr>
<td>60,000</td>
<td>94.52223</td>
<td>94.70715</td>
</tr>
</tbody>
</table>

If we look at table 3, we can notice also the fact that when sample sizes are relatively small, the accuracy of RBFNs is also mostly better, and when sample sizes are relatively large, the accuracy of MLPs is better.

Experiments were also run using a medium-sized data set in the UCI machine learning repository called ‘statlog’. The data set consists of the multi-spectral values of pixels in 3 by 3 neighborhoods in a satellite image and the classification associated with the central pixel in each neighborhood. In the data set, the class of a pixel is coded as a number, and there are seven classes, but there is no data for class 6. The total number of attributes is 36 which come from 4 spectral bands multiplied by 9 pixels in neighborhood, and all of them have numerical values in the range 0 to 255. The total number of instances is 6,435.

The statlog data set has relatively small number of instances compared to the other three data sets and all attributes are numeric. So, 6 were chosen as parameter value for clusters for K-means clustering. In order to train MLPs the given number of hidden layers is eighteen.

Table 4 shows the result of training for the two neural network algorithms. The initial sample size for training is 400, and the size of samples is doubled as the while loop runs, and we stop sampling when the sample size reaches to about half of the data set size. The rest of the data set after sampling is used for testing, so we have bigger test set data when sample size is small.

Table 4. RBFNs and MLPs for ‘statlog’ data set with different sizes of training data set

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Accuracy of RBFN(%)</th>
<th>Accuracy of MLP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>84.62718</td>
<td>84.58163</td>
</tr>
<tr>
<td>800</td>
<td>85.11978</td>
<td>86.4463</td>
</tr>
<tr>
<td>1,600</td>
<td>87.26538</td>
<td>87.17753</td>
</tr>
<tr>
<td>3,200</td>
<td>87.49035</td>
<td>88.49035</td>
</tr>
</tbody>
</table>

If we look at table 4, we can notice the fact that there is almost no relationship between sample size and accuracy between the two neural networks. So, more experiments were done for some middle sample sizes. Table 5 shows the result of training at some middle sample sizes that were not considered at the experiment in table 6. It also shows that when sample sizes are relatively small, the accuracy of RBFNs is mostly better, and when sample sizes are relatively large, the accuracy of MLPs is better.

Table 5. RBFNs and MLPs for ‘statlog’ data set with other different sizes of training data set

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Accuracy of RBFN(%)</th>
<th>Accuracy of MLP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>600</td>
<td>84.9621</td>
<td>84.84213</td>
</tr>
<tr>
<td>2,400</td>
<td>87.33123</td>
<td>88.11173</td>
</tr>
<tr>
<td>2,800</td>
<td>87.46305</td>
<td>88.23328</td>
</tr>
</tbody>
</table>

5 Conclusion
There are many successful neural network algorithms for data mining or machine learning. Among many artificial neural networks multilayer perceptrons and radial basis function networks are two representative neural network algorithms that are widely used for classification. Many researchers reported that the performance of radial basis function networks are better than that of multilayer perceptrons for their applications, but some other researchers reported the opposite results. This somewhat conflicting reports may be due to the fact that the performance of the algorithms dependent on the property of available training data set as well as the training data set size.

We experimented RBF networks and MLPs for classification tasks to see the effect of training data set change. A repeated sampling method with various sample sizes is applied to find out if there is any dependency between data set size and the performance of the neural network algorithms. Depending on data sets, RBF
networks and MLPs show different accuracies and the accuracy difference do not change as the training data set size grows. But in some data sets the accuracy of MLPs become better and better, then the accuracies of MLPs become better than those of RBF networks, even though those of MLPs are worse than those of RBF networks at first. As a conclusion, we can say that we should at least check the accuracy of neural networks in several sample sets, possibly with different sizes. The conclusion was drawn by experiments with four real world data sets.

References: