A comparison of RBF networks and random forest in forecasting ozone day

Hyontai Sug

***Abstract*—** It is known that random forest has good performance for data sets containing some irrelevant features, and it is also known that the performance of random forest is very good at ozone day prediction data set that is supposed to have some irrelevant features. On the other hand, it is known that when data sets do not contain irrelevant features, RBF networks are good at prediction tasks. Moreover, in general, we do not have exact knowledge about irrelevant features, because data space is usually far greater than available data for training. So we want to test that the two facts are true or not for the ozone data set. Experiments were done with random forests and RBF networks using k-means clustering, and showed that RBF networks are slightly better than random forest for the ozone day prediction.

***Keywords*—** RBF networks, random forest, decision trees, irrelevant features.

I. Introduction

Neural networks and decision trees are widely accepted for classification tasks in data mining or machine learning, and because each knowledge model has its own characteristic, finding appropriate knowledge models with the smallest error rates for given data sets is crucial for the success of data mining tasks [1], [2], [3], [4]. Even though the two knowledge models are the most successful data mining or machine learning methodologies, there are some weak points for each method because of the fact that they are built based on greedy algorithms and usually by the knowledge of experts.

Radial basis function (RBF) networks belong to one of major neural networks, and draw many researchers' attention because of good performance in many application fields [5], [6], [7].

Radial basis function makes an approximation based on training data, and Gaussian function is used mostly as the radial basis function [8], [9]. In order to train RBF networks first we should find appropriate centre and radius of radial basis function. For this task, we may use some unsupervised learning algorithms like k-means clustering, because k-means clustering algorithm is one of the mostly used algorithm for clustering [10].

Even though decision trees are widely accepted for data mining or machine learning tasks, they have some weak points like data fragmentation. So, sometimes decision trees have relatively poor accuracy compared to other knowledge models like neural networks. In order to overcome the problem, a large number of decision tree are generated for the same data set, and used simultaneously for prediction. Random forest [11], [12], [13] is one of such method, and known to be robust for irrelevant features with very good performance. So, random forest algorithm is applied to some data sets like ozone data set that is guessed to have some irrelevant features by domain experts [14], and the random forest algorithm showed very good result. In this paper, we want to compare the performance of RBF networks and random forest especially for the ozone data set, because we do not have exact knowledge about irrelevant features in the data set, and moreover, depending on data sets RBF networks are known to have very good performance.

In section 2, we provide the related work to our research, and in sections 3 we present some detail about random forest and RBF networks, and in section 4 we present our method of experiment. Experiments were run to see the effect of the method in section 5. Finally section 6 provides some conclusions.

II. Related Work

There is a big difference in training time between neural networks and decision trees. Generally, it takes far longer time to train neural networks than decision trees. But the two knowledge models are used very widely, because each one has its own good points. There are two kinds of networks based on how the networks are interconnected – feed-forward neural networks and recurrent neural networks [15]. RBF networks are one of the most popular feed-forward networks [16]. The training time of RBF networks is relatively shorter than other neural network algorithms. A good point of RBF networks is their good prediction accuracy with small-sized data sets, which is also true for other neural networks.

When we have very large data sets for training, we may use decision tree algorithms to save training time. There have been a lot of efforts to build better decision trees with respect to accuracy. C4.5 [17] is a fast and dirty type algorithm that was developed in early 90's, and is often referred in literature because of its wide availability [18]. Random forest [11], [19] uses many decision trees simultaneously for prediction so that it can avoid the negative effect of irrelevant features. A good point of decision tree algorithms is their scalability so that they

This work was supported by Dongseo University, "Dongseo Frontier Project" Research Fund of 2010.

are also good for very large data sets. There are scalable decision tree algorithms for large data sets like SLIQ [20], SPRINT [21], and PUBLIC [22]. SLIQ saves computing time especially for continuous attributes by using a pre-sorting technique in tree-growth phase, and SPRINT is an improved version of SLIQ to solve the scalability problem by building trees in parallel. PUBLIC tries to save some computing time by integrating the steps of pruning and generating branches. In [23] the authors compared the performance of four different neural networks, backpropagation network, RBF network, fuzzy-ARTUP-Net, LVQ, with binary and n-ary decision trees in industrial radiographic testing data, and showed the superiority of the four neural networks. On the contrary, Zang and Fan [14], [24] showed that bagging decision trees or random forest is the best predictors for ozone day prediction in their experiment. But they omitted some possible performance comparison with other neural networks. So we want to see some other alternative data mining method like RBF networks could generate better prediction accuracy for the data sets empirically.

III. RANDOM FOREST AND RBF NETWORK

We apply two existing data mining algorithms; radial basis function networks, and random forest. Random forest consists of many decision trees, and each tree votes for a class based on its own classification result. Each tree is constructed using the following algorithm:

1. F: the number of features to choose randomly.
2. N: the number of training examples.
3. Choose a training set of size N randomly by choosing N times without replacement.
4. Generate a decision tree based on F with no pruning.

The used decision tree algorithm for random forest is CART [25]. Parameters for random forest are the number of trees in the forest, and the number of features to choose randomly, F. According to Breiman [11], the number of trees in the forest can be 100, and F can be the first integer less than $\log_2 K + 1$, where K is the number of features of the target data set. But, because Zang and Fan [14], [24] recommended 30 as the number of trees to generate and the total number of features (72) of the ozone data set as the value for F, we also use the values for our experiment.

The task of forecasting with RBF network is a classification or regression problem, so the problem can be stated as a function approximation problem. Center point and radius are two parameters for Gaussian radial basis function. The center of the radial basis function indicates the central position, and the radius determines how the function spreads around its center. When we use Gaussian as a basis function, mean is the center and variance is the radius.

In order to train RBF networks first we should find appropriate center and radius of radial basis function. For this task, we may use some unsupervised learning algorithms like k-means clustering. After deciding the centers and radiuses logistic regression can be used to predict a class. So, K, the number of clusters in RBF network, is an important parameter that we can choose.

IV. THE METHOD OF EXPERIMENT

We used four random sample sets of size 200, 400, 600, 800, 1,000, 1,200, 1,400, 1,600 to see the trend of accuracy change with the two algorithms. For each random data set three different random forests are generated, and a decision tree of C4.5 is generated for reference.

We use RBF network that is based on k-means clustering, and because we want to find the best one, we increase the number of clusters incrementally, until some predefined limit. The following is a brief description of the procedure to find the best RBF network.

1. Initialize the_number_of_clusters as two;
2. Generate RBF network with the_number_of_clusters;
/* the accuracy of the RBF network is the base accuracy$_0$ */
3. best_accuracy := base_accuracy$_0$;
4. **Repeat** m times
   4.1 the_number_of_clusters := the_number_of_clusters + two;
   4.2 Generate RBF network with the_number_of_clusters;
   4.3 **If** the accuracy of the RBF network is greater than the best_accuracy **Then** best_accuracy := the accuracy of the RBF network;
End Repeat

In the algorithm depending on the size of available training data set, we set the value of m appropriately, and the number of clusters is incremented by the number of classes, which is two. In the experiment below m is set to larger values, if the size of training data set is larger.

V. EXPERIMENTATION

Experiments were run using data sets in UCI machine learning repository [26] called 'ozone'. The number of instances in ozone data set is 2,536. The data set consists of two different data sets – one hour and eight hour data set. The total number of features or attributes is 73, and one of them is class attribute having two classes.

The data set has large number of attributes compared to the available data, and many attributes have missing values also. Zang and Fan [14],[24] guessed that there might be some irrelevant attributes, so that they preferred random forest to single decision tree to average the effect of the irrelevant attributes.

We used RBF network using k-means clustering to train for a variety number of clusters and also used random forest using CART [27] for comparison. The following table 1 to table 18

shows the result of experiments. Table 1 thru 9 represent results for ozone one hour data set, and Table 10 thru 18 represent results for ozone eight hour data set. For experiment C4.5 uses default parameter values.

The number in the parentheses after the accuracy of the RBF network in the tables is the number of clusters. The parameters for the three random forests are given differently; for random forest 1 (RF1) 30 trees with 72 features, for random forest 2 (RF2) 100 trees with 7 features, for random forest 3 (RF3) 30 trees with 7 features.

The parameter values of 30 trees and 72 features are based on suggested values by Zang anf Fan's paper [14], [24]. The parameter values of 100 trees with 7 features are based on Breiman's [11]. The parameter values of 30 trees and 7 features are a combination of Zang and Fan's with Breiman's. Finally, 'RF avg' in the tables is the average accuracy of three random forests. All accuracies are represented in percentage.

A. Experiments for Ozone One Hour data Set

For the experiments for sample size 200 and 400 of ozone one hour data set in table 1 and 2, m was initialized with 10.

Table 1-1. results for 'ozone one hour' data set for sample size 200

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	96.6182	97.0462	94.7774	96.0616
RBFN	97.1318 (2)	97.0462 (2)	97.1318 (2)	97.1318 (2)
RF1	96.7466	96.9178	97.0462	96.8322
RF2	97.1318	97.0462	97.1318	97.1318
RF3	96.8322	97.0462	97.1318	97.1318
RF avg	96.9035	97.0034	97.1033	97.0219

Table 1-2. average accuracy for 'ozone one hour' data set for sample size 200

	Average accuracy
C4.5	96.1259
RBFN	97.1104 (2)
RF1	96.8857
RF2	97.1104
RF3	97.0355
RF avg	97.0153

If we look at table 1-1, we can notice that the RBFNs have the same accuracy with the best accuracy of random forest for each random sample set. If we look at table 1-2, we can notice that the average accuracy of RBFNs is the same with the best average accuracy of the random forests.

Table 2-1. results for 'ozone one hour' data set for sample size 400

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	95.412	94.6161	97.0037	97.0519
RBFN	97.191	97.191	97.0037	97.0519
	(2)	(2)	(2)	(2)
RF1	97.3315	97.1442	97.0037	97.0051
RF2	97.1442	97.191	97.0037	97.0519
RF3	97.1442	97.191	97.0037	97.0051
RF avg	97.2067	97.1754	97.0037	97.0207

Table 2-2. average accuracy for 'ozone one hour' data set for sample size 400

	Average accuracy
C4.5	96.0209
RBFN	97.1094 (2)
RF1	97.1211
RF2	97.0977
RF3	97.086
RF avg	97.1016

If we look at table 2-1, we can notice that the RBFNs have the same accuracy with the best accuracy of random forest for random sample set 2, 3, and 4. But for random sample set 1, random forest 1 has better accuracy than that of RBFN. If we look at table 2-2, we can notice that the average accuracy of RBFNs is slightly smaller than the best of random forests.

Table 3-1. results for 'ozone one hour' data set for sample size 600

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	95.8161	95.5579	96.2829	96.6942
RBFN	97.2107 (2)	97.2624 (2)	97.0057 (4)	97.5207 (2)
RF1	97.0558	97.2107	96.9541	97.3657
RF2	97.2107	97.2624	96.9541	97.469
RF3	97.2107	97.2624	96.9541	97.5207
RF avg	97.1591	97.2452	96.9541	97.4518

Table 3-2. average accuracy for 'ozone one hour' data set for sample size 600

	Average accuracy
C4.5	96.0878
RBFN	97.2499 (2.5)
RF1	97.1466
RF2	97.2241
RF3	97.2370
RF avg	97.2025

If we look at table 3-1, we can notice that the RBFNs have the same accuracy with the best accuracy of random forest for each random sample set 1, 2, and 4. But for random sample set 3, the RBFN has the best accuracy. If we look at table 3-2, we can notice that the average accuracy of RBFNs is slightly better than that of random forests.

For the experiments of sample size 800, 1,000, 1,200, 1,400, and 1,600 in table 4 to 8, m was initialized with 20.

Table 4-1. results for 'ozone one hour' data set for sample size 800

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	97.3502	96.3155	95.6211	96.1406
RBFN	97.235 (2)	97.1215 (2)	97.235 (2)	97.0046 (6)
RF1	97.235	97.1215	97.235	96.371
RF2	97.235	97.1215	97.235	96.371
RF3	97.235	97.1215	97.235	96.7742
RF avg	97.235	97.1215	97.235	96.5054

Table 4-2. average accuracy for 'ozone one hour' data set for sample size 800

	Average accuracy
C4.5	96.3569
RBFN	97.1490 (3)
RF1	97.1318
RF2	96.9906
RF3	97.0914
RF avg	97.0242

If we look at table 4-1, we can notice that the RBFNs have the same accuracy with the best accuracy of random forest for each random sample set 1, 2, and 3. But for random sample set 4, the RBFN has the best accuracy. If we look at table 4-2, we can notice that the average accuracy of RBFNs is slightly better than that of random forests.

Table 5-1. results for 'ozone one hour' data set for sample size 1,000

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	95.3776	97.1373	96.4216	97.0703
RBFN	97.3307 (2)	97.1373 (2)	96.7469 (2)	97.0703 (2)
RF1	97.3307	97.1373	96.7469	97.0703
RF2	97.3307	97.1373	96.7469	97.0703
RF3	97.3307	97.1373	96.7469	97.0703
RF avg	97.3307	97.1373	96.7469	97.0703

Table 5-2. average accuracy for 'ozone one hour' data set for sample size 1,000

	Average accuracy
C4.5	96.5017
RBFN	97.0713 (2)
RF1	97.0713
RF2	97.0713
RF3	97.0713
RF avg	97.0713

If we look at table 5-1, we can notice that the RBFNs have the same accuracy with the best accuracy of random forest for each random sample set, and we also can notice that C4.5 has good performance. Table 5-2 shows the result in average accuracy, so we can notice that there is no difference in accuracy between RBFNs and random forests for sample size 1,000.

Table 6-1. results for 'ozone one hour' data set for sample size 1,200

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	96.1826	95.6619	97.083	95.7535
RBFN	97.006 (26)	97.1578 (2)	97.5318 (12)	96.9311 (42)
RF1	97.006	97.1578	97.3822	96.8563
RF2	96.9311	97.1578	97.5318	96.7814
RF3	96.9311	97.1578	97.5318	96.7814
RF avg	96.9561	97.1578	97.4819	96.8064

Table 6-2. average accuracy for 'ozone one hour' data set for sample size 1, 200

	Average accuracy
C4.5	96.1703
RBFN	97.1767 (20.5)
RF1	97.1006
RF2	97.1005
RF3	97.1005
RF avg	97.1005

If we look at table 6-1, we can notice that the RBFNs have the same accuracy with the best accuracy of random forest for random sample sets 1, 2, and 3. But for sample set 4, the RBFN has the best accuracy. If we look at table 6-2, we can notice that the average accuracy of RBFNs is slightly better than that of random forests.

Table 7-1. results for 'ozone one hour' data set for sample size 1,400

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	95.6866	96.6549	96.9517	95.7784
RBFN	96.6549 (34)	97.2711 (40)	97.1856 (2)	97.2735 (2)
RF1	96.4789	97.0951	97.1856	97.2735
RF2	96.4789	97.0951	97.1856	97.2735
RF3	96.4789	97.0951	97.1856	97.2735
RF avg	96.4789	97.0951	97.1856	97.2735

Table 7-2. average accuracy for 'ozone one hour' data set for sample size 1,400

	Average accuracy
C4.5	96.2679
RBFN	97.0963(19.5)
RF1	97.0083
RF2	97.0083
RF3	97.0083
RF avg	97.0083

If we look at table 7-1, we can notice that the RBFNs have

the same accuracy with the best accuracy of random forest for random sample sets 3 and 4. But for sample sets 1 and 2, the RBFN has the best accuracy. If we look at table 7-2, we can notice that the average accuracy of RBFNs is slightly better than that of random forests.

Table 8-1. results for 'ozone one hour' data set for sample size 1,600

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	97.8306	96.3714	95.5128	95.2991
RBFN	97.624 (26)	97.0117 (30)	96.688 (18)	97.1154 (10)
RF1	96.5812	96.7983	96.5812	96.7949
RF2	96.9017	96.7983	96.5812	97.0085
RF3	96.9017	96.7983	96.5812	97.1154
RF avg	96.7949	96.7983	96.5812	96.9729

Table 8-2. average accuracy for 'ozone one hour' data set for sample size 1,600

	Average accuracy
C4.5	96.2535
RBFN	97.1098(21)
RF1	96.6889
RF2	96.8224
RF3	96.8492
RF avg	96.7868

If we look at table 8-1, we can notice that the RBFNs have the same accuracy with the best accuracy of random forest for random sample set 4. But for sample sets 1, 2 and 3, the RBFN has the best accuracy. If we look at table 8-2, we can notice that the average accuracy of RBFNs is slightly better than that of random forests.

All in all for ozone one hour data set, we can infer that RBFN is better than random forest as the summery in table 9.

Table 9. comparison of the best accuracy of RBFN and random forest based on sample size for ozone one hour data sets

Sample size	RBFN	Random forest
200	Same	Same
400		Better
600	Better	
800	Better	
1,000	Same	Same
1,200	Better	
1,400	Better	
1,600	Better	

B. Experiments for Ozone Eight Hour Data Set

Next we present the result of experiment for ozone eight hour data set. For the experiments for sample size 200 and 400 of ozone eight hour data set in table 10 and 11, m was initialized with 10.

Table 10-1. results for 'ozone eight hour' data set for sample size 200

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	92.0443	92.2879	89.9703	90.3171
RBFN	93.7901 (2)	93.7446 (2)	93.9227 (2)	93.7446 (2)
RF1	93.9186	93.916	94.0501	93.8732
RF2	94.0043	93.7446	93.9652	93.7446
RF3	93.9186	93.7875	93.9652	93.7875
RF avg	93.9472	93.8160	97.9935	93.8018

Table 10-2. average accuracy for 'ozone eight hour' data set for sample size 200

	Average accuracy
C4.5	91.1549
RBFN	93.8005 (2)
RF1	93.9395
RF2	93.8647
RF3	93.8647
RF avg	93.8896

If we look at table 10-1, we can notice that the RBFNs have the same accuracy with the worst accuracy of random forest for each random sample set. If we look at table 10-2, we can notice that the average accuracy of RBFNs is slightly smaller than that of random forests.

Table 11-1. results for 'ozone eight hour' data set for sample size 400

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	91.0497	90.2062	92.1747	90.956
RBFN	94.0019 (2)	93.4864 (2)	93.7207 (2)	93.8144 (2)
RF1	94.3768	93.5801	93.8613	93.8144
RF2	94.0956	93.5333	93.7207	93.8144
RF3	94.0487	93.3927	93.8144	93.8613
RF avg	94.1737	93.5020	93.7988	93.8300

Table 11-2. average accuracy for 'ozone eight hour' data set for sample size 400

	Average accuracy
C4.5	91.0967
RBFN	93.7559 (2)
RF1	93.9082
RF2	93.791
RF3	93.7793
RF avg	93.8261

If we look at table 11-1, we can notice that the RBFNs have the same accuracy with the worst accuracy of random forest for random sample set 3 and 4. But for random sample set 1, all random forests have better accuracy than that of RBFN, and for random sample set 2, the accuracy of RBFN is middle among

the accuracies of random forests. If we look at table 11-2, we can notice that the average accuracy of RBFNs is slightly smaller than that of random forests.

For the experiments of sample size 600, 800, 1,000, 1,200 in table 12 to 15, m was initialized with 20.

Table 12-1. results for 'ozone eight hour' data set for sample size 600

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	91.1582	90.5377	92.3475	92.6557
RBFN	92.9162 (2)	93.8469 (2)	93.5884 (2)	93.8469 (2)
RF1	93.0196	94.0538	93.8987	94.1055
RF2	93.0196	93.8987	93.6401	93.8987
RF3	93.0196	94.0021	93.5884	93.8987
RF avg	93.0196	93.9849	93.7091	93.9676

Table 12-2. average accuracy for 'ozone eight hour' data set for sample size 600

	Average accuracy
C4.5	91.6748
RBFN	93.5496 (2)
RF1	93.7694
RF2	93.6143
RF3	93.6272
RF avg	93.6703

If we look at table 12-1, we can notice that the RBFNs have some poorer accuracy than random forests. If we look at table 12-2, we can notice that the average accuracy of RBFNs is slightly smaller than that of random forests.

Table 13-1. results for 'ozone eight hour' data set for sample size 800

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	89.9077	92.2722	91.4648	92.6182
RBFN	93.887 (2)	94.223 (36)	93.4409 (34)	93.5409 (2)
RF1	93.9446	93.4833	93.7716	93.8293
RF2	93.8293	93.714	93.4833	93.5986
RF3	93.8293	93.7716	93.5986	93.5409
RF avg	93.8677	93.6563	93.6178	93.6563

Table 13-2. average accuracy for 'ozone eight hour' data set for sample size 800

	Average accuracy
C4.5	91.5657
RBFN	93.773 (18.5)
RF1	93.7572
RF2	93.6563
RF3	93.6851
RF avg	93.6995

If we look at table 13-1, we can notice that the RBFN has the best accuracy for sample set 2, and the secondly best accuracy for sample set 1. But for random sample sets 3 and 4, the RBFN have inferior accuracy. On the other hand, if we look at table 13-2, we can notice that the average accuracy of RBFNs is slightly better than that of random forests.

Table 14-1. results for 'ozone eight hour' data set for sample size 1,000

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	92.5949	92.029	91.8293	91.7683
RBFN	94.4308 (10)	94.2029 (2)	94.0244 (2)	93.9024 (20)
RF1	94.5532	94.5048	94.3293	93.5976
RF2	94.2472	94.3237	94.0854	93.4756
RF3	94.3084	94.3237	94.0244	93.4756
RF avg	94.3696	94.3841	94.1464	93.5163

Table 14-2. average accuracy for 'ozone eight hour' data set for sample size 1,000

	Average accuracy
C4.5	92.0554
RBFN	94.1401 (8.5)
RF1	94.2462
RF2	94.033
RF3	94.033
RF avg	94.1041

If we look at table 14-1, we can notice that the RBFNs have some better accuracy values for random sample set 4, and inferior accuracy values for sample sets 1, 2 and 3. Table 14-2 shows the result in average accuracy, so we can notice that RBFN is slightly inferior to random forests for sample size 1,000.

Table 15-1. results for 'ozone eight hour' data set for sample size 1,200

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	92.5787	91.979	91.4543	92.4288
RBFN	93.5532 (2)	94.2279 (2)	93.3283 (2)	92.9535 (16)
RF1	94.078	94.5277	93.6282	93.1034
RF2	93.5532	94.5277	93.3283	93.0285
RF3	93.6282	94.3028	93.3283	92.9535
RF avg	93.7531	94.4527	93.4283	93.0285

Table 15-2. average accuracy for 'ozone eight hour' data set for sample size 1, 200

	Average accuracy
C4.5	92.1102
RBFN	93.5157(5.5)
RF1	93.8343
RF2	93.6094
RF3	93.5532
RF avg	93.6657

If we look at table 15-1, we can notice that the accuracies of RBFNs are slightly inferior to the accuracies of random forests. If we look at table 15-2, we can notice that the average accuracy of RBFNs is slightly inferior to that of random forests.

For the experiments of sample size 1,400, and 1,600 in table 16 and 17, m was initialized with 40.

Table 16-1. results for 'ozone eight hour' data set for sample size 1,400

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	92.328	93.1278	93.1278	92.9515
RBFN	94.4444 (20)	94.2731 (2)	94.2731 (18)	92.9515 (20)
RF1	94.4444	94.7137	94.2731	92.1278
RF2	94.0035	94.3612	94.3612	92.8634
RF3	94.0035	94.3612	94.4493	92.8634
RF avg	94.1505	94.4787	94.36122	92.6182

Table 16-2. average accuracy for 'ozone eight hour' data set for sample size 1,400

	Average accuracy
C4.5	92.8838
RBFN	93.9855(15)
RF1	93.8898
RF2	93.8973
RF3	93.9194
RF avg	93.9021

If we look at table 16-1, we can notice that the RBFNs have the best accuracy of random forest for random sample sets 1 and 4. But for sample sets 2 and 3, the random forests have the best accuracy. If we look at table 16-2, we can notice that the average accuracy of RBFNs is slightly better than that of random forests.

Table 17-1. results for 'ozone eight hour' data set for sample size 1,600

	Sample 1	Sample 2	Sample 3	Sample 4
C4.5	92.5054	92.6203	93.262	91.7559
RBFN	94.0043 (70)	93.9037 (40)	94.2446 (70)	94.2184 (22)
RF1	94.3255	93.5829	94.0107	94.4325
RF2	93.8972	93.0481	93.5829	93.8972
RF3	94.0043	93.0481	93.5829	94.0043
RF avg	94.0757	93.2264	93.7255	94.1113

Table 17-2. average accuracy for 'ozone eight hour' data set for sample size 1,600

	Average accuracy
C4.5	92.5329
RBFN	94.0928(50.5)
RF1	94.0879
RF2	93.6034
RF3	93.6599
RF avg	93.7401

If we look at table 17-1, we can notice that the RBFNs have the best accuracies for random sample sets 2 and 3. But for sample sets 1 and 4, random forest 1 has the best accuracy. If we look at table 17-2, we can notice that the average accuracy of RBFNs is slightly better than that of random forests.

All in all, we can summery as in table 18 for the ozone eight hour data set.

Table 18. comparison of the best accuracy of RBFN and random forest based on sample size for ozone eight hour data sets

Sample size	RBFN	Random forest
200		Better
400		Better
600		Better
800	Better	
1,000		Better
1,200		Better
1,400	Better	
1,600	Better	

So, we can see that the accuracy of RBF network becomes better as the sample size grows. If we consider both data sets of ozone one hour and eight hour data sets, because RBF network is better in 5 cases for ozone one hour data set, and it is better in 3 cases for ozone eight data set, but random forest is better in 1 case for ozone one hour data set, and it is better in 5 cases for ozone eight data set, we can conclude that RBF network is slightly better than random forest for ozone data set.

VI. CONCLUSIONS

Radial basis function (RBF) networks are widely accepted for data mining or machine learning tasks in which available data set size is relatively small. Moreover, when the data sets do not include many irrelevant features, it is known that RBF networks are one of the most successful data mining or machine learning tools for classification. But, RBF 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 some critical parameters are defined by the knowledge of experts. So, some improvements may be possible.

Because most RBF networks use clustering algorithms, we need to set appropriate number of clusters for best accuracy. But, determining the appropriate number of clusters is arbitrary in nature, so we incremented the number of clusters progressively to find some better RBF networks of accuracy in systematic manner, especially for ozone data set that is known to be best predicted by ensemble of decision tree-based method.

Even though the ozone data set might contain some irrelevant attributes, by applying RBF network to the data set repeatedly with varying number of clusters, we found that RBF network is slightly superior to random forest of decision trees in

accuracy. Especially RBF network is better for ozone one hour data set, and it is better when training data set size is relatively larger for ozone eight hour data set. Experiment with several sample sizes showed the trend.

REFERENCES

[1] R.J. Roiger, M.W. Geatz, *Data Mining: A Tutorial-Based Primer*, Addison Wesley, 2003.

[2] Z. Zainuddin, O. Pauline, "Function Approximation Using Artificial Neural Networks", *WSEAS Transactions on Mathematics*, vol. 7, issue 6, 2008, pp. 333-338.

[3] K. Komiya, C. Igarashi, K. Shibahara, K. Hojimoto, Y. Tasima, Y. Kotani, "Generating a Set of Rules to Determine the Gender of a Speaker of a Japanese Sentence", *WSEAS Transactions on Communications*, vol. 8, issue 1, 2009, pp. 112-121.

[4] M. Yu, L. Chen, "A Diachronic Study of POP Album Cover From 1980 to 1992 in Taiwan", *WSEAS Transactions on Communications*, vol. 8, issue 10, 2009, pp. 1064-1075.

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

[6] J. Heaton, *Introduction to Neural Networks for C#*, 2nd ed., Heaton Research Inc., 2008.

[7] G. Baylor, E.I. Konukseven, A.B. Koku, "Control of a Differentially Driven Mobile Robot Using Radial Basis Function Based Neural Networks", *WSEAS Transactions on Systems and Control*, vol. 3, issue 12, 2008, pp. 1002-1013.

[8] R.P. Lippmann, "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, vol. 3, no. 4, 1987, pp. 4-22.

[9] R.J. Howlett, L.C. Jain, *Radial Basis Function Networks I: recent developments in theory and applications*, Physics-Verlag, 2001.

[10] S. Russel, P. Novig, *Artificial Intelligence: a Modern Approach*, 2nd ed., Prentice Hall, 2002.

[11] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, 2001, pp. 5-32.

[12] J.D. Rebolledo-Mendez, M. Higashihara, Y. Yamada, K. Satou, "Characterization and Clustering of GO Terms by Feature Importance Vectors Obtained from Microarray Data", *WSEAS Transactions on Biology and Biomedicine*, vol. 5, issue 7, 2008, pp. 163-172.

[13] H. He, C. Che, F. Ma, J. Zhang, X. Luo, "Traffic Classification Using En-semble Learning and Co-Training", *8th WSEAS International Conference on Applied Informatics and Communications*, 2008, pp. 458-463.

[14] K. Zhang, W. Fan, "Forecasting skewed biased stochastic ozone days: analyses, solutions and beyond", *Knowledge and Information Systems*, vol.14, no. 3, pp. 299-326, 2008.

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

[16] M.J.L. Orr, *Introduction to Radial Basis Function Networks*, http://www.anc.ed.ac.uk/~mjo/intro.ps, 1996.

[17] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, Inc., 1993.

[18] D.T. Larose, *Data Mining Methods and Models*, Wiley-Interscience, 2006.

[19] W. Fan, H. Wang, P.S. Yu, S. Ma, "Is random model better? On its accuracy and efficiency", *Proceedings of third IEEE International Conference on Data Mining (ICDM2003)*, 2003.

[20] M. Mehta, R. Agrawal, and J. Rissanen, "SLIQ : A Fast Scalable Classifier for Data Mining", *EDBT'96*, Avignon, France, 1996.

[21] . Shafer, R. Agrawal, and M. Mehta., "SPRINT : A Scalable Parallel Classifier for Data Mining", *Proc. 1996 Int. Conf. Very Large Data Bases*, Bombay, India, 1996, pp. 544 – 555.

[22] R. Rastogi, K. Shim, "PUBLIC : A Decision Tree Classifier that Integrates Building and Pruning", *Data Mining and Knowledge Discovery*, Vol. 4, No. 4, Kluwer International, 2002, pp. 315 – 344.

[23] P. Perner, U. Zscherpel, C. Zacobsen, "A comparison between neural networks and decision trees based on data from industrial radiographic testing", *Pattern Recognition Letters*, 2001, pp. 47-54.

[24] K. Zhang, W. Fan, X. Yuan, I. Davidson, X. Li, "Forecasting Skewed Biased Stochastic Ozone Days: Analyses and Solutions", *6th International Conference on Data Mining*, 2006, pp. 753-764.

[25] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth International Group, 1984.

[26] A. Suncion, D.J. Newman, *UCI Machine Learning Repository* [http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html]. Irvine, CA: University of California, School of Information and Computer Sciences, 2007.

[27] I.H. Witten, E. Frank, *Data Mining*, 2nd ed., Morgan Kaufmann, 2005.

Hyontai Sug received the B.S. degree in Computer Science and Statistics from Busan National University, Busan, Korea, in 1983, the M.S. degree in Computer Science from Hankuk University of Foreign Studies, Seoul, Korea, in 1986, and the Ph.D. degree in Computer and Information Science & Engineering from University of Florida, Gainesville, FL, in 1998. He is an associate professor of the Division of Computer and Information Engineering of Dongseo University, Busan, Korea from 2001. From 1999 to 2001, he was a full time lecturer of Pusan University of Foreign Studies, Busan, Korea. He was a researcher of Agency for Defense Development, Korea from 1986 to 1992. His areas of research include data mining and database applications.