International Journal of Fuzzy Logic and Intelligent Systems, vol. 9, no. 1, March 2009 pp. 36-41

Improvement of SOM using Stratification

Sunghae Jun

Department of Bioinformatics & Statistics, Cheongju University, 360-764 Chungbuk, Korea

Abstract

Self organizing map(SOM) is one of the unsupervised methods based on the competitive learning. Many clustering works have been performed using SOM. It has offered the data visualization according to its result. The visualized result has been used for decision process of descriptive data mining as exploratory data analysis. In this paper we propose improvement of SOM using stratified sampling of statistics. The stratification leads to improve the performance of SOM. To verify improvement of our study, we make comparative experiments using the data sets form UCI machine learning repository and simulation data.

Key Words : Self Organizing Map(SOM), Data Visualization, Clustering, Stratified Sampling

1. Introduction

Self organizing map(SOM) is an artificial neural network for unsupervised learning by Kohonen[1]. Also SOM is a visualization method in the clustering algorithms. It has two layers which are input layer and output layer(feature map). This clustering algorithm chooses the winner node in feature map by minimum of the Euclidean norm distance from input vector[2]. Most clustering methods, such as K-means clustering algorithm and hierarchical clustering methods, are needed the number of clusters[1],[3]. Their clustering results are depended on the numbers[4],[5],[6]. For example, we get improper clustering results by K-means clustering algorithm when wrong number of clusters is used. But the researcher may not know the correct number of clusters in the given data[7]. In general we can know the number of clusters roughly using SOM[2]. The feature map shows the visualization of the data points. The similar points are mapped to same node of feature map. So we are able to determine the number of clusters using the results of feature map. In this paper, we propose improved SOM using stratified sampling. Our research contributes on an objective determination of the number of clusters in the feature map of SOM by proposed contribution rate measure and stratified sampling theory of statistics. We verify improved performance of our work by experimental results using data sets from UCI machine learning repository and simulation.

2. Related Researches

There are many researches about the improvements of SOM[3],[8],[9],[10]. Recently probabilistic approaches have been considered to improve the clustering results of SOM[18],[23],[24]. The statistical and probabilistic theory support SOM with theoretic base[11],[12],[13]. We use stratified sampling as a statistical method for improving SOM. In statistical sampling theory, stratification is to decide the population into some strata[14]. Then each stratum is not similar to other strata. After stratification we perform simple random sampling from the strata[10]. Many studies of machine learning have used this approach of efficient data reduction[15],[16],[17],[18]. In this paper, we use 50% stratified sampling to improve SOM heuristically.

3. Stratification for SOM

We apply stratification to total given training data set. That is, we consider stratified sampling to reduce the data set for improving the clustering results of SOM. In this paper, we use not simple random sampling but stratified sampling. The simple random sampling does not include the information of the classes in data. The information is very important to cluster analysis. To solve this problem, we use the stratified sampling in our research[14]. Before doing SOM, we divide total data into strata which are n disjoint classes, ST_1, ST_2, ..., ST_n. The sum of all classes, (ST_1+ST_2+...+ST_n) is equal to total data. Each stratum is shown as the following.

$$(ST\_i, \#\text{ of points in } ST\_i), \quad i=1,2,\ldots,n \qquad (1)$$

To get the reduce training samples from the strata, we do simple random sampling in each stratum ST_i. A sample i(SA_i) is formed from ST_i. So the sample data with SA_1, SA_2, ..., SA_n are constructed by sampling from ST_1, ST_2, ..., ST_n.

We perform learning SOM in the data set which has SA_1, SA_2, ..., SA_n. This data set is the reduced data from given total training data. We can get more visualized feature map of

Manuscript received Dec. 30, 2008; revised Mar. 10, 2009.

SOM because the reduced training data are based on stratification. The following figure shows the proposed SOM clustering and the general SOM clustering.

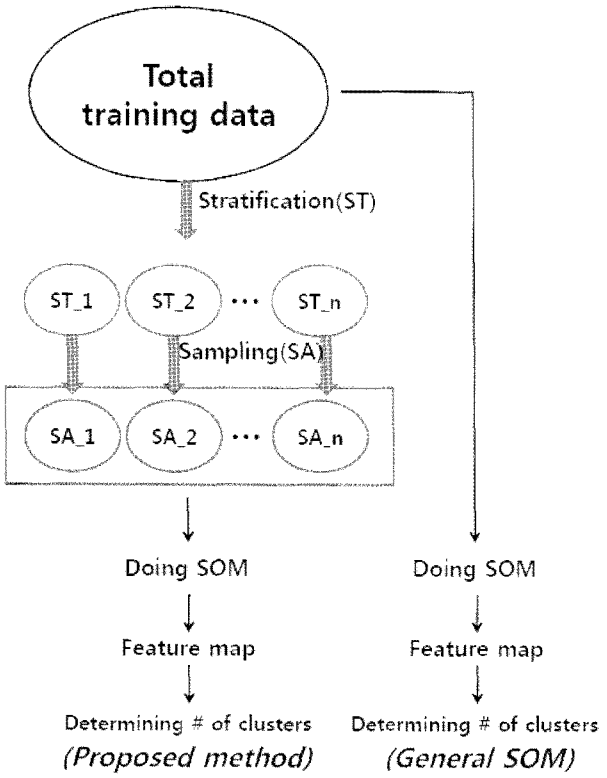

Fig. 1 Proposed method & general SOM

In the above figure, our proposed SOM method has the stratified sampling process. But the general SOM method has not the stratification. The main goal of our research is the improvement of SOM using the stratified sampling method of statistical sampling theory. This research contributes to the optimal number of clusters of the feature map in SOM. The map of SOM shows the clustering results of given training data. But we have a difficulty to determine the number of clusters because the feature map is not able to clear clustering result. By the stratified sampling we get the feature map more clearly. So we are able to determine the number of clusters effectively. To evaluate the improvement of our study we propose contribution rate(CR) value as a measurement as the following.

$$CR(\%)=(\#\text{ of used instances} / \#\text{ of total instances})\times 100 \quad (2)$$

In the above the # of used instances represents the number of used instances for determining the number of clusters. Also the # of total instances is the number of total instances in the given training data set. Larger CR value contributes to clearer determination of the number of clusters. So we are able to select the cluster size easily according as the CR value increases. We find the CR values of our method are larger than general SOM in our experimental results. The following shows the detailed explanation of our work.

(I) Given training data x_i , i=1,2,…,r with n classes
(II) Dividing total data into strata, ST_1, ST_2, …, ST_n
(III) Stratified sampling from strata to SA_1, SA_2, …, SA_n for given sampling rate
(IV) Doing SOM
  (IV-1) Initializing the weight vectors $w_j(0)$, j=1,2,…,s (s: the number of neurons)
  (IV-2) Computing similarity (selecting the winner neuron)
    $k(x)=\arg\min_j\|x(m)-w_j\|$
    where k(x) is the winner neuron at time step m.
  (IV-3) Updating weights
    $w_j(m+1)=w_j(m)+\alpha(m)(x(m)-w_j(m))$
    where α is learning rate.
  (IV-4) Stopping SOM until satisfying given conditions
(V) Assigning all instances to fixed clusters

Next we make experiments to verify improved performances of our method. The analytical tool of the experiment is the R project for statistical computing[19]. Also we use the 'som' and 'sampling' packages of R project.

4. Experiments and Results

To verify our improved performances of stratified sampling based SOM by CR value, we use the data sets from UCI machine learning repository and simulation data[20],[21],[22]. Firstly we make an experiment using data from UCI machine learning repository. The following table shows the summary information of training data set.

Table 1. Data Summary from UCI Machine Learning

Data Sets	# of Instances	# of Attributes (continuous)
Ozone Level Detection	2536	73
Poker Hand	2510	10
Thyroid Domain	3772	6

In above, three data sets are Ozone level detection, Poker hand, and Thyroid domain data. We find the number of instances and the number of continuous attributes in table 1. In this paper, we use 50% stratified sampling on each data set. So, we compare total training data and 50% stratified sample by the CR. The following figures show the results of total and sample data sets by SOM. First two figures are the clustering results of feature map of SOM by Ozone level detection data. Figure 2 is the feature map of SOM by total training data. This figure shows two major nodes in feature map of SOM. They are the nodes with n=1270 and n=960. Where n is the number of instances. So

we know the number of clusters is two in Ozone level detection data. Also the CR value of this result is computed by the following.

$$[(1270+960) / (1270+166+140+960)]\times 100 = 87.93(\%) \quad (3)$$

The other experiments of our paper are performed by the figure and CR value.

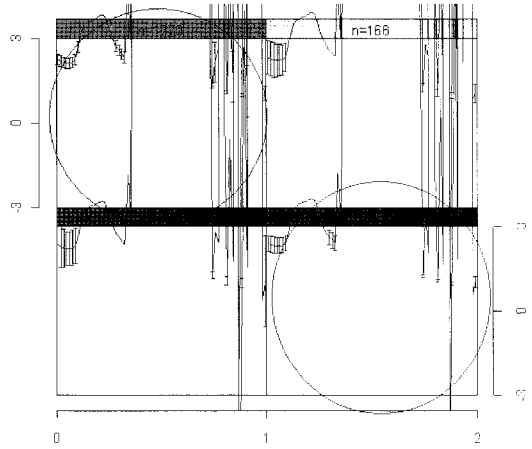

Fig. 2 Ozone level detection (total data)

Also we find the clustering result of feature map of SOM using the 50% stratified sampling data set in figure 3.

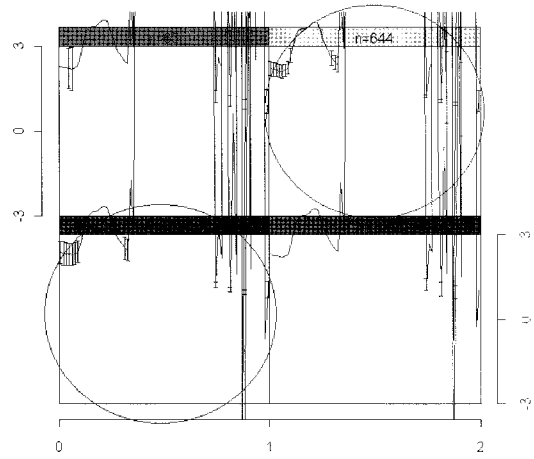

Fig. 3 Ozone level detection (50% sampling data)

We can compute the CR value of Ozone level detection data by 50% stratified sampling from figure 3. Next figure 4 and figure 5 show the clustering results of SOM with total and 50% sampling data sets. The CR values can be got by the figures same as the CR values of Ozone level detection data were computed.

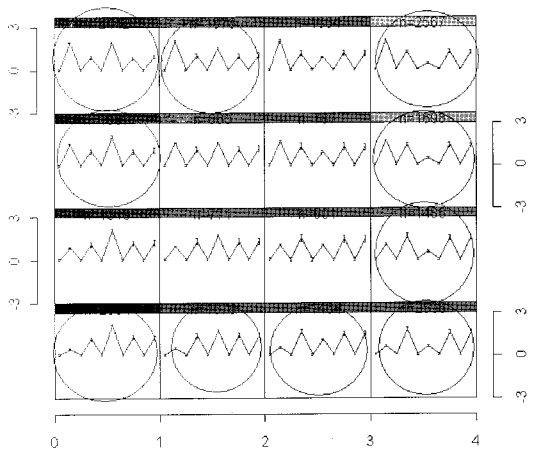

Fig. 4 Poker hand (total data)

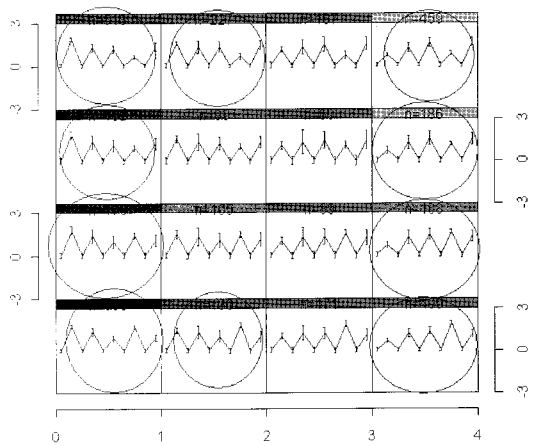

Fig. 5 Poker hand (50% sampling data)

The following two figures show the results of feature maps from the Thyroid domain data with total and 50% sampling.

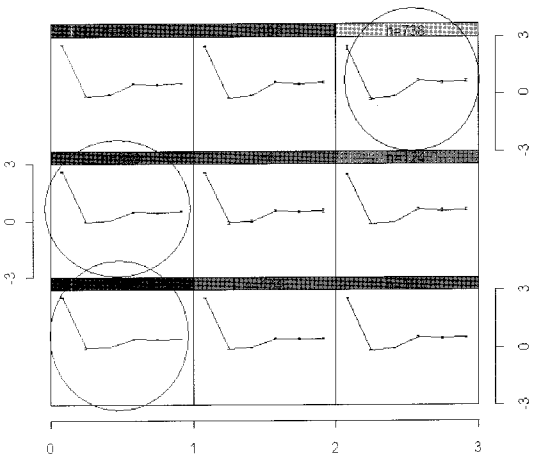

Fig. 6 Thyroid Domain (total data)

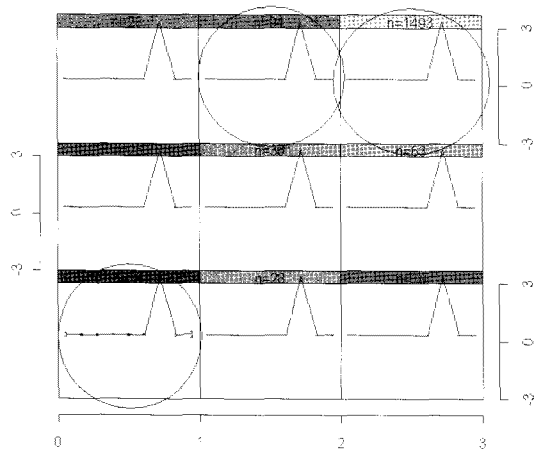

Fig. 7 Thyroid Domain (50% sampling data)

We know the CR values of the clustering results by total and 50% stratified sampling from the UCI machine learning repository in the following table.

Table 2. Clustering Results of Total and Stratified Sampling (Data from UCI Machine Learning Repository)

Data Sets	# of Clusters	Sampling	CR value(%)
Ozone Level Detection	2	Total	87.93
		50% Strata	90.15
Poker Hand	10	Total	77.32
		50% Strata	81.67
Thyroid Domain	3	Total	74.28
		50% Strata	88.87

In the above table we verify the improved performance of proposed stratified sampling based SOM. The CR values of 50% stratified sampling data are smaller than the values of total data. To verify the performances by simulation we use the data sets as the following table.

The simulation data have three types according to the number of clusters. Table 3 shows the parameters of our simulation data. We generate the data from normal distribution with the parameter values of mean and standard deviation.

The following figures from figure 8 to figure 13 show the clustering results of SOM according to total and 50% stratified sampling data sets.

Table 3. Parameters of Simulation Data

# of Clusters	Group Index	Normal Distribution	
		Mean	S.D.
2	1	0.0	0.3
	2	0.5	0.6
5	1	0.0	0.3
	2	0.5	0.6
	3	1.0	0.9
	4	1.5	1.2
	5	2.0	1.5
10	1	0.0	0.2
	2	0.5	0.4
	3	1.0	0.6
	4	1.5	0.8
	5	2.0	1.0
	6	2.5	1.2
	7	3.0	1.4
	8	3.5	1.6
	9	4.0	1.8
	10	4.5	2.0

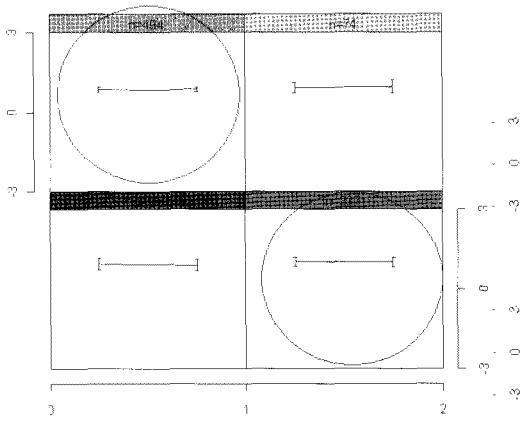

Fig. 8 Simulation: # of clusters=2 (total data)

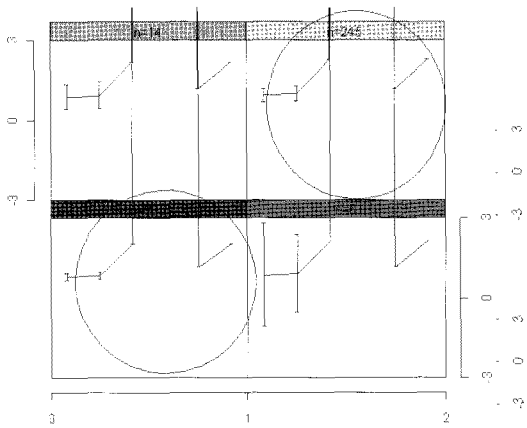

Fig. 9 Simulation: # of clusters=2 (50% sampling data)

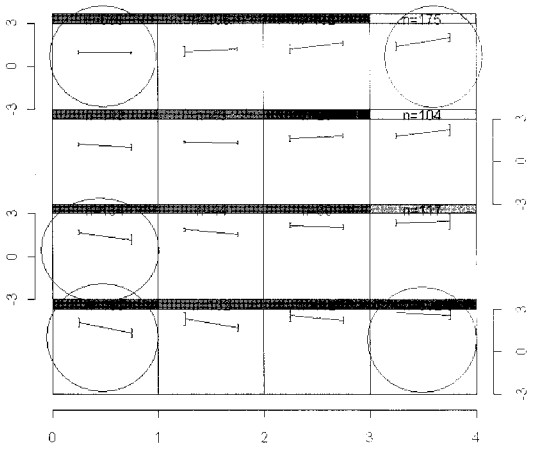

Fig. 10 Simulation: # of clusters=5 (total data)

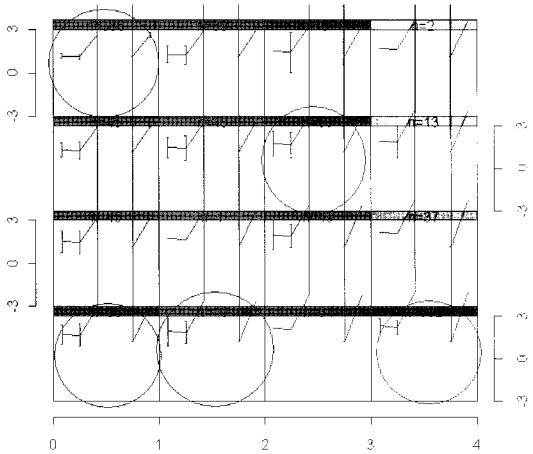

Fig. 11 Simulation: # of clusters=5 (50% sampling data)

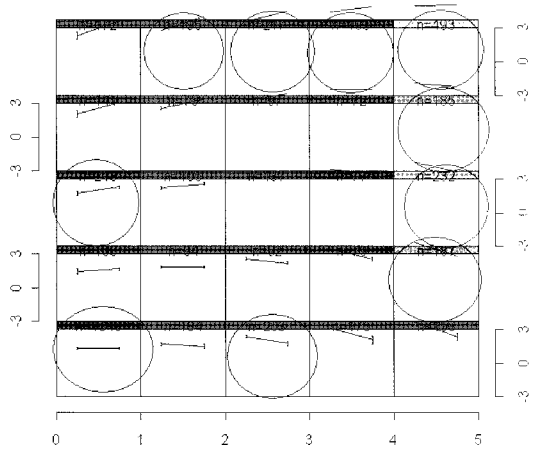

Fig. 12 Simulation: # of clusters=10 (total data)

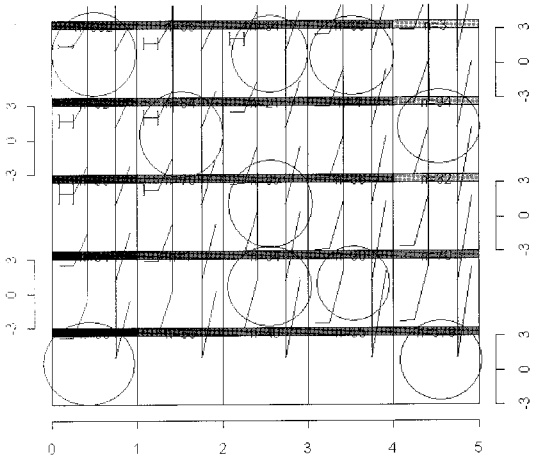

Fig. 13 Simulation: # of clusters=10 (50% sampling data)

We find the improved performances of proposed SOM by the following table.

Table 4. Clustering Results of Total and Stratified Sampling (Simulation Data)

# of Clusters	# of Instances	Sampling	Contribution Rate(%)
2	1000	Total	83.10
		50% Strata	96.00
5	2500	Total	61.60
		50% Strata	75.04
10	5000	Total	63.28
		50% Strata	75.92

The CR values of 50% stratified sampling based SOM are smaller than the values of general SOM. So we can verified the contribution of our proposed SOM.

5. Conclusions and Future Works

In this paper we showed improvement of SOM using stratification. The stratification is a sampling theory of statistics using strata. We use the stratified sampling to improve the performance of general SOM in our research. For the verification of the work we also proposed the CR measure in the feature map of SOM.

To construct the improved classification and regression models we consider the stratification. These will be our future works.

References

[1] B. S. Everitt, S. Landau, M. Leese, *Cluster Analysis*, Arnold, 2001.

[2] T. Kohonen, *Self Organizing Maps*, Second Edition, Springer, 1997.

[3] J. Han, M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2001.

[4] S. H. Jun, "An Optimal Clustering using Hybrid Self Organizing Map," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 6, no. 1, pp. 10-14, 2006.

[5] S. H. Jun, "New Heuristic of Self Organizing Map using Updating Distribution," *Advances in Cognitive Neurodynamics*, Book Chapter 170, Springer, 2008.

[6] D. A. Stacey, R. Farshad, "A probabilistic self-organizing classification neural network architecture," *Proceeding of International Joint Conference on Neural Networks*, vol. 6, pp. 4059-4063, 1999.

[7] P. Giudici, *Applied Data Mining*, Wiley, 2003.

[8] A. L. N. Fred, A. K. Jain, "Robust Data Clustering," *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 128-133, 2003.

[9] C. M. Bishop, M. Svensen, C. K. I. Williams, "GTM: A Principled Alternative to the Self Organizing Map," *Proceeding of ICANN 1996*, vol. 1112, pp. 165-170, 1996.

[10] A. Ngan, S. Thiria, F. Badran, M. Yaccoub, C. Moulin, M. Crepon, "Clustering and classification based on expert knowledge propagation using probabilistic self-organizing map(PRSOM): application to the classification of satellite ocean color TOA observations," *Proceeding of IEEE International Symposium on Computational Intelligence for Measurement Systems and Applications*, pp. 146-148, 2003.

[11] A. Utsugi, "Topology selection for self-organizing maps," *Network: Computation in Neural Systems*, vol. 7, no. 4, pp. 727-740, 1996.

[12] A. Utsugi, "Hyperparameter selection for self-organizing maps," *Neural Computation*, vol. 9, no. 3, pp. 623-635, 1997.

[13] S.-H. Jun, "Improvement of Self Organizing Maps using Gap Statistic and Probability Distribution," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 8 no. 2, pp. 116-120, 2008.

[14] S.K. Thompson, *Sampling*, 2nd ed., John Wiley & Sons, 2002, pp. 117-127.

[15] C.S. Ding, Q. Wu, C.T. Hsieh, and M. Pedram, "Stratified Random Sampling for Power Estimation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 6, pp. 465-471, 1998.

[16] P.A.D.I. Santos, Jr., R.J. Burke, and J.M. Tien, "Prograssive Random Sampling With Stratification," *IEEE Transactions on Systems, Man, and Cybernetics*, part A, vol. 37, no. 6, pp. 1223-1230, 2007.

[17] M. Xing, M. Jaeger, and H. Baogang, "An Effective Stratified Sampling Scheme for Environment Maps with Median Cut Method," *Proceedings of International Conference on Computer Graphics, Imaging and Visualisation*, pp. 384-389, 2006.

[18] M. Keramat, and R. Kielbasa, "A study of stratified sampling in variance reduction techniques for parametric yield estimation," *Proceedings of IEEE International Symposium on Circuits and Systems*, vol. 3, pp. 1652-1655, 1997.

[19] R Development Core Team, *R: A language and environment for statistical computing, R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2008,

[20] UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html

[21] W. L. Martinez, A. R. Martinez, *Computational Statistics Handbook with MATRAB*, Chapman & Hall, 2002.

[22] G. Mclachlan, D. Peel, *Finite Mixture Models*, John Wiley & Sons, 2000.

Sunghae Jun

He received the BS, MS, and PhD degrees in department of Statistics, Inha University, Korea, in 1993, 1996, and 2001. Also, He received PhD degree in department of Computer Science, Sogang University, Korea in 2007. He is currently Assistant Professor in department of Bioinformatics & Statistics, Cheongju University, Korea. He has researched statistical learning theory and evolutionary algorithms.

Phone : +82-43-229-8205
Fax : +82-43-229-8432
E-mail : shjun@cju.ac.kr