International Journal of Fuzzy Logic and Intelligent Systems, vol. 4, no. 1, June 2004 pp. 119-124

Sparse Data Cleaning using Multiple Imputations

Sung-Hae Jun*, Seung-Joo Lee* and Kyung-Whan Oh**

***Department of Statistics, Cheongju University**
360-764 Chungbuk, Korea
****Department of Computer Science, Sogang University**
121-742 Seoul, Korea

Abstract

Real data as web log file tend to be incomplete. But we have to find useful knowledge from these for optimal decision. In web log data, many useful things which are hyperlink information and web usages of connected users may be found. The size of web data is too huge to use for effective knowledge discovery. To make matters worse, they are very sparse. We overcome this sparse problem using Markov Chain Monte Carlo method as multiple imputations. This missing value imputation changes spare web data to complete. Our study may be a useful tool for discovering knowledge from data set with sparseness. The more sparseness of data in increased, the better performance of MCMC imputation is good. We verified our work by experiments using UCI machine learning repository data.

Key Words : Cleaning of Sparse data, Multiple Imputation, Markov Chain Monte Carlo

1. Introduction

The web log file contains a rich and dynamic collection of hyperlink information and web page access and usage information. It also seems to be so huge for effective data mining. The size of web log data is very large, but web log is very sparse. Many web pages in web server are not accessed by each user. So we have a difficulty for prediction modeling. It is very difficult to estimate dependency of web pages in spare web data. An efficient preprocessing approach is needed for this problem. Using the missing value imputation by multiple imputation method, the spare data are changed to perfect for prediction model. This imputation provides a useful strategy for dealing with data sets with spare. In this paper, we use MCMC(Markov Chain Monte Carlo) method for multiple imputation to replace missing data with estimated data. In our paper, the MCMC method was presented good prediction result in spare data. And we verified these results through experiments using UCI machine learning repository data set[15].

2. Necessity of Sparse Data Cleaning

The sparsity of data as web log file is made by several reasons[3],[4]. For example, it occurs when all pages of web server are more than user visited web pages. This is frequent case in web log data. Therefore, the click stream data of cleaned web log file is very sparse. Generally, this sparsity is extreme. So, we have a difficulty of web log analysis, for example, web usage mining with web information recommendation, next web page prediction, and web page duration time forecasting. The web data with sparsity is not analyzed by general methods, for example, regression, multi-layer perceptron(MLP) and others[5]. The MCMC method as missing value imputation is very useful tool for sparsity data analysis. In this paper, the elimination of sparse from sparse data is performed using MCMC as multiple imputation method.

Manuscrip received Mar. 22, 2004; revised Apr. 30, 2004.

3. Cleaning using Missing Value Imputation

3.1 Multiple imputation

Missing data is a problem in a data sets and frequently complicates data analysis for scientific investigation. The development of statistical methods for dealing with data sets with missing values has been an active area of research in recent decades. Imputation is a general method for handling missing data problem. A strategy is single imputation. This method can be applied to impute a single value for each missing value. Traditional approaches include case deletion and mean imputation. In the last decade the main interest has centered on regression imputation and imputation using EM(Expectation-Maximization) algorithm. Instead of filling in a single value for each missing value, Rubin proposed multiple imputation which replaces each missing value with a set of plausible values by drawing from the conditional distribution of the missing data given the observed data[10],[11]. Multiple imputation is the method of choice for complex incomplete data problems. Some methods apply only to special missing data patterns, where others apply to any other pattern.

Assume that the missing data is missing at random(MAR).

With a monotone missing data pattern, simple methods has been proposed, including regression method, prospensity methods and predictive mean matching for continuous variables[8],[9],[11]. With an arbitrary or general missing data pattern, MCMC methods has been suggested. The regression, predictive mean matching and MCMC methods require assumptions that the data are from a multivariate normal distribution, but there is some evidence that the inferences tend to be robust to minor departures from this assumption[13].

We briefly review MCMC method since in general the web log data sets has arbitrary missing patterns. In Bayesian inference, the posterior probability distribution contains all the current information about the unknown parameters. Using Bayes's theorem, the posterior distribution of parameters θ is computed by

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \quad (1)$$

where $p(y|\theta)$ is often called the likelihood function of θ given y and $p(\theta)$ is called the prior distribution of θ.

MCMC has been applied as a method for exploring complicated posterior distribution in Bayesian inference. Let Y denote the $n \times p$ matrix of complete data. Denote the observed part of Y by Y_{obs}, and the missing part by Y_{mis}, so that $Y=(Y_{obs}, Y_{mis})$. Data augmentation algorithm is applied to Bayesian inference with missing data by repeating the following two steps.

Imputation I-step:

With the estimated mean vector and covariance matrix, I-step first draws a value of the missing data from a conditional predictive distribution Y_{mis} given Y_{obs},

$$Y_{mis}^{(t+1)} \sim p(Y_{mis}|Y_{obs}, \theta^{(t)}) \quad (2)$$

Posterior P-step:

The P-step draw a new value of θ from the complete data posterior,

$$\theta^{(t+1)} \sim p(\theta|Y_{obs}, Y_{mis}^{(t+1)}) \quad (3)$$

Repeating I-step and P-step from a starting value $\theta^{(0)}$, this create a Markov chain

$$(Y_{mis}^{(1)}, \theta^{(1)}), (Y_{mis}^{(2)}, \theta^{(2)}), \cdots \quad (4)$$

which converges in stationary distribution $p(Y_{mis}, \theta|Y_{obs})$. Schafer in [13] called (2) Imputation or I-step and (3) the Posterior or P-step.

In this study, the posterior mode with a noninformative prior was computed from the EM algorithm and was used as the starting value from the chain. After the completion of m imputation, we computed the MSE.

3.2 Other imputation algorithm

There are many methods in the imputation algorithm. In our research, we compared the some imputation methods with MCMC imputation. The following are the comparative methods with proposed MCMC method. These are many used in data mining tools as SAS E-Miner[16].

Tree imputation: Generally, when missing data take place in continuous cases, the mean or conditional mean imputation methods are used. But these imputation methods have the low predictive accuracy. To overcome the problem tree imputation model which is nonparametric approach is considered[6]. Suppose the data set is defined by $(y_{1i}, \cdots, y_{ji}, \cdots, y_{pi}, t)$, $(i=1,2,\cdots,n)$, where p is the dimension of input vector, n is the size of objects, and t is a target variable. A tree construction arises from a divide and conquer algorithmic strategy which recursively divides the data space into two subregions according to a splitting criterion which aims to optimize classification or prediction for the cases to be split[1]. The replacement values of tree imputation are estimated by analyzing each input as a target, and the remaining input and rejected variables as predictors. Variables with a model role of target cannot be used to impute the data. Because the imputed value for each input variable is based on the other input variables, this imputation technique may be more accurate than simply using the variable mean or conditional mean to replace the missing values[2],[6].

Distribution-based imputation: The replacement values of distribution-based imputation are calculated based on the random percentiles of the variable's distribution. In this case, values are assigned based on the probability distribution of the non-missing observations. This imputation method typically does not change the distribution of the data very much[6].

Robust M-Estimators of Location: Tukey's biweight, Hubers, and Andrew's wave are robust M-estimators of location. Common estimators such as the sum of squared residuals can become unstable when using outlier data points and distort the resulting estimators. M-Estimators try to reduce the effect of outliers by using substitute functions which are symmetric, have a unique minimum at zero, and increase less than standard squared residuals under summation. There are a wide variety of M-estimators. A suitably chosen M-estimator will have the two properties. One is robustness of efficiency in larger samples. The other is resistance to outliers or gross errors in the data. An estimator has robustness of efficiency over a range of distributions if its variance is close to the minimum for each distribution. Robustness of efficiency guarantees that the estimator is good when repeated samples are drawn from a distribution that is not known precisely. An estimator is resistant if it is not changed much by small groups of outliers or by rounding and grouping errors among observations. Robustness of efficiency and resistance are the main reason why you would want to use one of the M-estimators for imputation. The default tuning constant for each M-estimator is as follows [6].

Table 1. The default training constant for each M-estimator

Estimator	Default tuning constant
Tukey's biweight	9
Huber	1.5
Andrew's wave	6.283185

4. Experimental Results

In this section, we want to show the experimental results of proposed MCMC by ABALONE data set from UCI machine learning repository[15]. The number of instances of data is 4177. The 8 attributes which are length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, and rings are abalone's physical state. The abalone data is complete. For our experiments, we make complete abalone data to incomplete. The incomplete abalone data have 5%, 10%, 20%, 30%, 40%, 50%, and 60% missing ratios. Currently, since the tree imputations have been good preprocessing methods of missing data, we compared the MCMC multiple imputation with the tree imputations and other imputation methods. Compared with these imputations, the MCMC imputation method was better. Our verified results are shown in following tables and figures. In our experiment, the MSE of each table is computed as following[14].

$$MSE_i = \frac{1}{n}\sum_{j=1}^{n_i}(y_{ij} - y^*_{ij})^2 \quad (5)$$

and

$$MSE = \frac{1}{m}\sum_{i=1}^{m} MSE_i \quad (6)$$

where y_{ij} represents the jth known value of ith variable, y^*_{ij} represents the jth predicted value of ith variable and n_i is the number of missing data for ith variable. The m is the number of imputation. The Mean of each method in the table is the average of all variable's MSE.

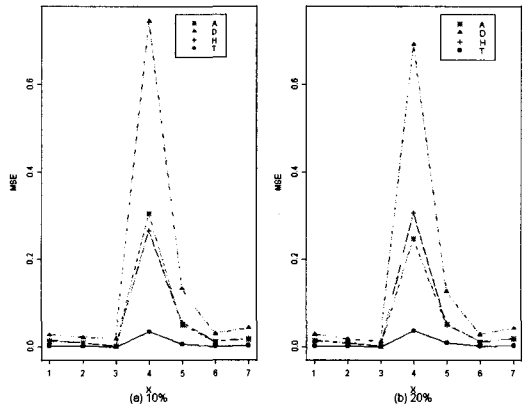


Fig. 1. MSE plots of comparative methods

The MSEs of comparative methods which are Andrew's wave, Huber, distribution based method, and tree imputations are shown in Table 2. Also Fig. 1 shows the MSE plots of comparative methods about 10% and 20% missing rates. In this figure, A, D, H, and T represent Andrew's wave, Huber, distribution based method, and tree imputations, respectively. From Table 2 and Fig. 1, we know that the tree imputation has better performance than others. This MSE is lower than other methods. In variable 4, the difference of MSEs among 4 methods is shown large. This is because the variance of variable 4 is larger than others. We also found that the values of MSE are increased, as the rates of missing increased. Following table shows MSE of MCMC method as multiple imputation.

Table 2. MSE of the comparative methods

Method	Y	Proportion of missing		
		5%	10%	20%
Andrew's wave	Y1	0.0140	0.0132	0.0145
	Y2	0.0097	0.0094	0.0088
	Y3	0.0015	0.0016	0.0014
	Y4	0.2656	0.3047	0.2467
	Y5	0.0564	0.0508	0.0516
	Y6	0.0118	0.0118	0.0119
	Y7	0.0188	0.0179	0.0186
	Mean	0.0540	0.0585	0.0505
Huber	Y1	0.0304	0.0141	0.0133
	Y2	0.0191	0.0097	0.0094
	Y3	0.0133	0.0015	0.0016
	Y4	0.6208	0.2656	0.3062
	Y5	0.1474	0.0566	0.0509
	Y6	0.0348	0.0119	0.0119
	Y7	0.0502	0.0188	0.0180
	Mean	0.1309	0.0540	0.0588
D. Based	Y1	0.0333	0.0270	0.0278
	Y2	0.0176	0.0211	0.0182
	Y3	0.0082	0.0184	0.0117
	Y4	0.7126	0.7472	0.6937
	Y5	0.1336	0.1339	0.1269
	Y6	0.0329	0.0296	0.0281
	Y7	0.0512	0.0443	0.0433
	Mean	0.1413	0.1459	0.1357
Tree	Y1	0.0015	0.0014	0.0023
	Y2	0.0008	0.0014	0.0019
	Y3	0.0003	0.0004	0.0006
	Y4	0.0369	0.0344	0.0361
	Y5	0.0051	0.0059	0.0091
	Y6	0.0013	0.0015	0.0018
	Y7	0.0018	0.0027	0.0030
	Mean	0.0068	0.0068	0.0078

In Table 3, m is the finite number of imputation data set. For example, if m is 3 we replace each missing cell with 3 predictive values. The MSE plots of MCMC methods on 5%, 10%, 20%, 30%, 40%, and 50% are shown in Fig. 2, 3 and 4. The MSE values of MCMC methods are very small. Also we know that all MSE values from variable 1 to variable 7 of MCMC methods are smaller than comparative methods.

We used another measure of performance which is mean absolute deviation(MAD). MAD is the average of the absolute difference between the original(target) value and the predicted value. This measure can be expressed as

$$MAD_i = \frac{1}{n_i}\sum_{j=1}^{n_i}|y_{ij} - y^*_{ij}| \quad (7)$$

and

$$MAD = \frac{1}{m}\sum_{i=1}^{m} MAD_i \quad (8)$$

where y_{ij} represents the jth original value of ith variable, y^*_{ij} represents the jth predicted value of ith variable and n_i is the number of missing data for ith variable. The m is the number of imputation.

Table 3. MSE of the MCMC method

m	Y	Proportion of missing						
		5%	10%	20%	30%	40%	50%	60%
3	Y1	0.0006	0.0006	0.0009	0.0010	0.0014	0.0019	0.0021
	Y2	0.0004	0.0004	0.0007	0.0006	0.0010	0.0012	0.0014
	Y3	0.0004	0.0004	0.0005	0.0012	0.0010	0.0005	0.0006
	Y4	0.0033	0.0043	0.0059	0.0074	0.0121	0.0222	0.0226
	Y5	0.0022	0.0026	0.0046	0.0039	0.0060	0.0075	0.0081
	Y6	0.0010	0.0009	0.0012	0.0012	0.0015	0.0019	0.0020
	Y7	0.0012	0.0014	0.0016	0.0018	0.0024	0.0031	0.0033
	Mean	0.0013	0.0015	0.0022	0.0024	0.0036	0.0055	0.0057
5	Y1	0.0005	0.0005	0.0008	0.0008	0.0013	0.0017	0.0020
	Y2	0.0004	0.0004	0.0006	0.0006	0.0009	0.0011	0.0013
	Y3	0.0003	0.0003	0.0004	0.0012	0.0010	0.0005	0.0005
	Y4	0.0029	0.0041	0.0054	0.0068	0.0109	0.0212	0.0204
	Y5	0.0020	0.0025	0.0042	0.0036	0.0055	0.0069	0.0072
	Y6	0.0009	0.0008	0.0011	0.0011	0.0014	0.0017	0.0018
	Y7	0.0010	0.0013	0.0014	0.0016	0.0021	0.0029	0.0030
	Mean	0.0011	0.0014	0.0020	0.0022	0.0033	0.0051	0.0052
10	Y1	0.0004	0.0004	0.0007	0.0008	0.0011	0.0015	0.0018
	Y2	0.0004	0.0003	0.0006	0.0005	0.0008	0.0010	0.0012
	Y3	0.0003	0.0003	0.0004	0.0011	0.0010	0.0004	0.0004
	Y4	0.0026	0.0038	0.0050	0.0061	0.0096	0.0198	0.0190
	Y5	0.0018	0.0023	0.0039	0.0033	0.0049	0.0063	0.0067
	Y6	0.0008	0.0008	0.0010	0.0010	0.0012	0.0016	0.0017
	Y7	0.0009	0.0011	0.0013	0.0015	0.0019	0.0027	0.0028
	Mean	0.0010	0.0013	0.0018	0.0020	0.0029	0.0047	0.0048

Table 4 and Table 5 are shown the MADs of comparative methods and MCMC method, respectively. The Mean of each method in these tables represents the MAD, that is, the average of all variables's MAD. From these tables, we also found that the MADs of MCMC method are smaller than others. According to this result, the imputation method using MCMC approach has a good performance. We know that the cleaning performance of sparse data using the MCMC method is better than the general comparative methods. So, the MCMC method as multiple imputation can be used for cleaning sparse data which is web log file.

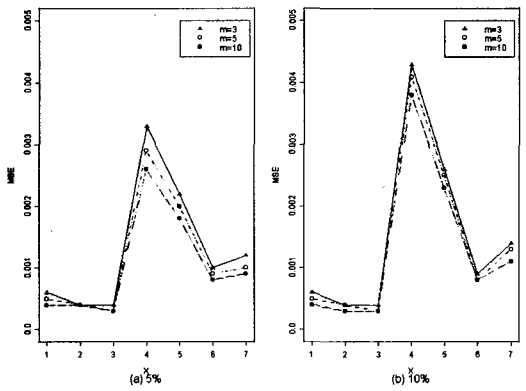


Fig. 2. MSE plots of MCMC methods: (5% and 10%)

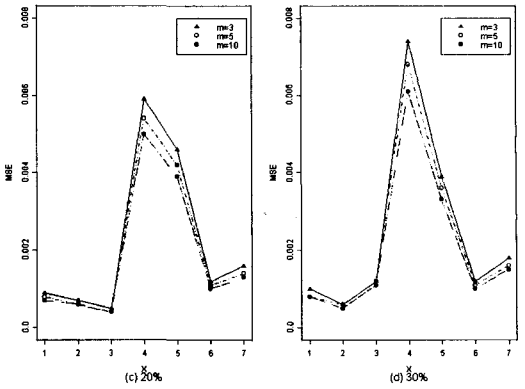


Fig. 3. MSE plots of MCMC methods:(20% and 30%)

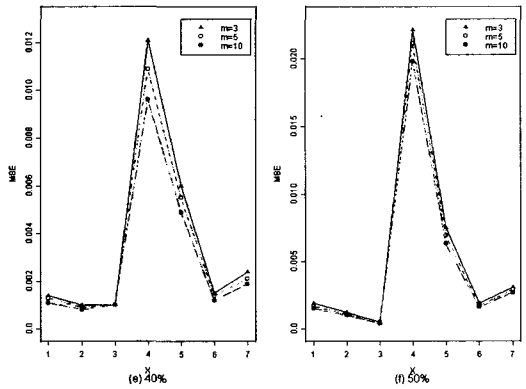


Fig. 4. MSE plots of MCMC methods:(40% and 50%)

Table 4. MAD of the comparative methods

Method	Y	Proportion of missing		
		5%	10%	20%
Andrew's wave	Y1	0.0934	0.0923	0.0982
	Y2	0.0785	0.0797	0.0758
	Y3	0.0310	0.0315	0.0312
	Y4	0.4296	0.4434	0.3884
	Y5	0.1924	0.1751	0.1882
	Y6	0.0869	0.0865	0.0882
	Y7	0.1118	0.1081	0.1127
	Mean	0.1462	0.1452	0.1404
Huber	Y1	0.0931	0.0923	0.0982
	Y2	0.0781	0.0792	0.0754
	Y3	0.0310	0.0317	0.0312
	Y4	0.4286	0.4436	0.3885
	Y5	0.1928	0.1749	0.1880
	Y6	0.0868	0.0864	0.0879
	Y7	0.1118	0.1082	0.1127
	Mean	0.1460	0.1452	0.1403
D. Based	Y1	0.1448	0.1341	0.1350
	Y2	0.1060	0.1147	0.1082
	Y3	0.0467	0.0592	0.0521
	Y4	0.6550	0.6888	0.6684
	Y5	0.2869	0.2852	0.2837
	Y6	0.1389	0.1355	0.1296
	Y7	0.1722	0.1611	0.1666
	Mean	0.2215	0.2255	0.2205

Table 4. MAD of the comparative methods

Method	Y	Proportion of missing		
		5%	10%	20%
Tree	Y1	0.0253	0.0291	0.0348
	Y2	0.0222	0.0260	0.0315
	Y3	0.0119	0.0143	0.0153
	Y4	0.1369	0.1323	0.1412
	Y5	0.0552	0.0553	0.0711
	Y6	0.0270	0.0282	0.0323
	Y7	0.0311	0.0365	0.0406
	Mean	0.0442	0.0460	0.0524

Table 5. MAD of the MCMC method

m	Y	Proportion of missing						
		5%	10%	20%	30%	40%	50%	60%
3	Y1	0.0182	0.0175	0.0216	0.0224	0.0265	0.0293	0.0329
	Y2	0.0142	0.0148	0.0181	0.0179	0.0214	0.0240	0.0259
	Y3	0.0155	0.0145	0.0155	0.0144	0.0143	0.0172	0.0190
	Y4	0.0364	0.0415	0.0496	0.0596	0.0717	0.0876	0.1004
	Y5	0.0318	0.0343	0.0429	0.0442	0.0527	0.0582	0.0628
	Y6	0.0223	0.0215	0.0250	0.0258	0.0276	0.0302	0.0324
	Y7	0.0237	0.0256	0.0276	0.0292	0.0335	0.0369	0.0404
	Mean	0.0232	0.0242	0.0286	0.0305	0.0354	0.0405	0.0448
5	Y1	0.0168	0.0166	0.0201	0.0210	0.0249	0.0275	0.0311
	Y2	0.0135	0.0140	0.0165	0.0167	0.0501	0.0225	0.0245
	Y3	0.0139	0.0133	0.0138	0.0138	0.0133	0.0156	0.0171
	Y4	0.0333	0.0390	0.0462	0.0547	0.0665	0.0823	0.0943
	Y5	0.0296	0.0329	0.0399	0.0416	0.0495	0.0544	0.0584
	Y6	0.0212	0.0215	0.0240	0.0241	0.0259	0.0282	0.0305
	Y7	0.0210	0.0237	0.0254	0.0275	0.0304	0.0341	0.0377
	Mean	0.0213	0.0230	0.0266	0.0285	0.0372	0.0378	0.0419
10	Y1	0.0158	0.0158	0.0188	0.0200	0.0236	0.0262	0.0296
	Y2	0.0128	0.0133	0.0156	0.0161	0.0190	0.0214	0.0235
	Y3	0.0126	0.0123	0.0126	0.0131	0.0126	0.0141	0.0153
	Y4	0.0306	0.0366	0.0438	0.0502	0.0612	0.0773	0.0892
	Y5	0.0267	0.0305	0.0378	0.0391	0.0462	0.0507	0.0555
	Y6	0.0188	0.0191	0.0219	0.0553	0.0243	0.0263	0.0290
	Y7	0.0196	0.0215	0.0237	0.0258	0.0280	0.0322	0.0350
	Mean	0.0196	0.0213	0.0249	0.0314	0.0307	0.0355	0.0396

5. Conclusion

In this paper, the MCMC imputation approach for sparseness elimination of very sparse data was proposed. This is based on multiple imputation theory. The advantage of this method is to change sparse data into complete. It is impossible that general preprocessing techniques can be used for sparse data cleaning. Our research will support knowledge discovery processes. Our future work is to develop the method of hybrid MCMC model to upgrade performance of sparse data cleaning.

Reference

[1] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, "*Classification and regression trees*", Wadsworth & Brooks, 1984.

[2] C. Conversano, C. Cappelli, "Missing data incremental imputation through tree based methods", *14th Conference on Computational Statistics*, 24-28 August 2002, Berlin, Germany, 2002.

[3] C. Cortes, V. Vapnik, "*Support Vector Networks*", Machine Learning, 1995.

[4] R. Fletcher, "*Practical Methods of Optimization*", John Wiley & Sons, Inc. New York, 1989.

[5] S. Haykin, "*Neural Networks*", 2nd edition. Prentice Hall, 1999.

[6] D. C. Hoaglin, F. Mosteller, J. W. Tukey, "*Understanding robust and exploratory data analysis*", John Wiley & Sons, Inc. New York, 1983.

[7] R. J. A. Lavori, R. Dawson, D. Shera, "*A Multiple Imputation Strategy for Clinical Trials with Truncation of Patent Data*", Statistics in Medicine, 1995.

[8] R. J. A. Little, "A Test of Missing Completely at Random for Multivariate Data with Missing Values", *Journal of the American Statistical Association*, 1988.

[9] P. R. Rosenbaum, D. B. Rubin, "*The Central Role of the Propensity Score in Observational Studies for Causal Effects*", Biometrica, 1983.

[10] D. B. Rubin, "*Inference with missing data*", Biometrika, 1976.

[11] D. B. Rubin, "*Multiple Imputation for Nonresponse in Surveys*", John Wiley & Sons Inc. New York, 1987.

[12] D. B. Rubin, "Multiple Imputation After 18+ Years", *Journal of the American Statistical Association*, 1996.

[13] J. L. Schafer, "*Analysis of Incomplete Multivariate Data*", Chapman and Hall. New York, 1997.

[14] V. Vapnik, "*The Nature of Statistical Learning Theory*", Springer. New York, 1995.

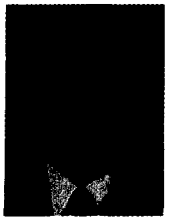
[15] http://www.ics.uci.edu/~mlearn/MLRepository.html

[16] http://www.sas.com

Sung-Hae Jun

Sung-Hae Jun received the B.S., M.S., and Ph.D. degrees in department of Statistics from Inha University, Korea, in 1993, 1996, and 2001. He is currently with the artificial intelligence laboratory in the department of computer science at Sogang University, Seoul, Korea, where he is a Ph.D. candidate. Also He is currently with the department of Statistics at Cheongju University, Chunkbuk, Korea, where he is full time lecturer. His research interests include artificial intelligence and data engineering.

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



Seung-Joo Lee

Seung-Joo Lee received the B.S. degree in department of Statistics from Cheongju University, Chungbuk, Korea, in 1985. Also he received the M.S., and Ph.D. degrees in department of Statistics from Dongkuk University, Seoul, Korea, in 1987, and 1995. He is currently with the department of statistics at Cheongju University, Chungbuk, Korea, where he is an associate professor. His research interests include Bayesian statistics and multi-variate analysis.

Phone : +82-43-229-8204
Fax : +82-43-229-8432
E-mail : access@cju.ac.kr

Kyung-Whan Oh

Kyung-Whan Oh received the B.S. degree in mathematics from Sogang University, Seoul, Korea, in 1978, and the M.S. and Ph.D. degrees in computer science from Florida State University, Tallahassee, in 1985 and 1988, respectively. He is currently with the department of computer science at Sogang University, Seoul, Korea, where he is a Professor. His research and teaching interests include fuzzy system, cognitive science, knowledge discovery and data mining, intelligent agents and multi-agent systems, expert system and statistical learning.

Phone : +82-2-703-7626
Fax : +82-2-704-8273
E-mail : kwoh@ccs.sogang.ac.kr