International Journal of Fuzzy Logic and Intelligent Systems, vol. 12, no. 1, March 2012, pp. 1-5
http://dx.doi.org/10.5391/IJFIS.2012.12.1.1 pISSN 1598-2645 eISSN 2093-744X

A Classification Method Using Data Reduction

Daiho Uhm[1], Sunghae Jun[2*] and Seung-Joo Lee[2]

[1]Department of Statistics, Oklahoma State University, Stillwater, OK 74078, USA
[2]Department of Statistics, Cheongju University, Chungbuk 360-764, Korea

Abstract

Data reduction has been used widely in data mining for convenient analysis. Principal component analysis (PCA) and factor analysis (FA) methods are popular techniques. The PCA and FA reduce the number of variables to avoid the curse of dimensionality. The curse of dimensionality is to increase the computing time exponentially in proportion to the number of variables. So, many methods have been published for dimension reduction. Also, data augmentation is another approach to analyze data efficiently. Support vector machine (SVM) algorithm is a representative technique for dimension augmentation. The SVM maps original data to a feature space with high dimension to get the optimal decision plane. Both data reduction and augmentation have been used to solve diverse problems in data analysis. In this paper, we compare the strengths and weaknesses of dimension reduction and augmentation for classification and propose a classification method using data reduction for classification. We will carry out experiments for comparative studies to verify the performance of this research.

Key words: Data reduction and augmentation, Gaussian mixture model, Principal component analysis, Support vector machine, K-nearest neighbor

1. Introduction

The dimension of data has been an important issue in data analysis [1-2]. Many researchers used dimension reduction and augmentation to solve problems in data analysis such as the curse of dimensionality and selection of optimal boundary [3-4]. Data reduction has been used in diverse fields of data analysis to overcome the curse of dimensionality [5-6]. Dimension reduction (DR) is deeply related to the curse of dimensionality. The curse of dimensionality is caused intractable problems of computing time in data analysis [3]. That is, the computing cost is one of the problems of the curse of dimensionality. As the number of variables increases, the computing cost increases exponentially. To avoid this problem, DR methods have been used [7]. We can analyze given data easily by reducing the dimension [8]. So, we have to develop the methods to reduce the number of original variables. Principal component analysis (PCA) and factor analysis (FA) methods are popular techniques in DR. The PCA and FA reduce the number of variables to avoid the curse of dimensionality. The curse of dimensionality is to increase the computing time exponentially in proportion to the number of variables. So, many methods have been published for dimension reduction [1],[3],[7-8]. Next, Data augmentation (DA) is another approach to efficient data analysis. All variables of original data are mapped to a feature space with high dimension in DA [7],[9]. On the contrary to DR, the DA augments the number of original variables [4]. This can solve some problems of data analysis such as non-separable in classification [10]. The DA approach was introduced after DR. But, recently many methods of DA have been published [4],[11]. Statistical learning theory (SLT) is a representative DA approach [11]. SLT has three versions of data analysis [2]. They are support vector machine (SVM), support vector regression (SVR), and support vector clustering (SVC) for classification, regression, and clustering respectively. The aim of this research is focused on classification. So, we will compare SVM with proposed method.

In previous research related to data dimension, we knew researchers were focused on one goal of DR or DA. There was not any research that considered DR and DA together. However, in this paper we consider DR and DA at the same time. We investigate the characteristics, strengths, and weaknesses of DR and DA. Also, we propose a classification method using DR based on Gaussian mixture. This research combines mixture model with K-nearest neighbor (*K-NN*) algorithm for an efficient classification. In general, Gaussian mixture model [12-13] was used for clustering task. However, we apply this model to classification task. We make experiments to compare the performances between some methods of DR and DA for classification. Furthermore, we compare the proposed method with traditional DR and DA methods using data sets from UCI machine learning repository [14].

2. Data Reduction and Augmentation

Since a serious problem of the curse of dimensionality is the burden of computing time as the number of variables increases [3],[8], many techniques of DR have been researched to avoid

Manuscript received Feb. 24, 2012; revised Mar. 5, 2012; accepted Mar. 7, 2012

*Corresponding Author: Sunghae Jun(shjun@cju.ac.kr)

the curse of dimensionality [3]. DR has two approaches as Figure 1.

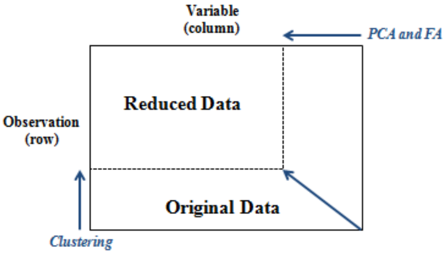

Fig. 1. Data reduction by PCA (FA) and clustering

First, we reduce the row dimension using clustering approach. This is to cluster all observations into groups. That is, clustering is row leveled DR approach. Second, PCA or FA is another DR approach based on columns. In many DR cases, PCA and FA are popular techniques for reducing the data dimension [2]. Also, these have been used for feature selection in data analysis [6-7],[15]. In this paper, we use PCA as a comparative DR method. It is because PCA is a useful approach to DR using data exploration [7]. PCA summarizes the original data set to a linear combination by new principal component (PC) [16]. The PC is represented as following:

$$PC = lx , \tag{1}$$

where the l and x are PC loading and input vector, respectively. Since the maximal number of PCs is d, where d is the number of columns in the original data, the number of PCs, m could be equal to d. However, we hope to have much smaller m than d ($m<<d$). Determining m depends on the proportion of total population variance [2]. The proportion (PR) of jth PC is shown as following [17]:

$$PR_{PC_j} = \frac{\lambda_j}{\lambda_1 + \lambda_2 + \ldots \lambda_m}, \tag{2}$$

where λ_k is the variance (proportion) of kth PC. This represents the explanation of kth PC for the data. In this paper, we select m, when the cumulative variance is to about 95%. Also, the minimum number of PCs is not less than two. A mapping from original the data space with d dimensions to the PC space with m dimensions is defined as following:

$$D_R(x) : R^d \rightarrow R^m . \tag{3}$$

$D_R(x)$ is a low-dimensional encoding of the original data x. For performing classification, we use PC score data as input data by $D_R(x)$. Classification is the process of constructing a classifier that determines data classes to predict the class of instances whose class is unknown [1]. Many classification algorithms have been researched in diverse fields [1]. They were based on statistics and machine learning. To compare the proposed method with DR and DA in classification field, we use *K-NN* as a classifier applied after our method and PC score. In *K-NN*, the instances are represented as objects in Euclidean space. Also, the nearest neighbor is defined by Euclidean distance [1]:

$$dist(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} , \tag{4}$$

where x_k is an object. A new object can be assigned to the most common class by its *K-NN* result. We control the size of the pattern space, *K*, by optimal classification.

Next, DA is to augment given data with high dimension. DA has also two approaches for augmenting data like as DR. These approaches are based on variables (columns) and observations (rows). Figure 2 shows general DA approaches.

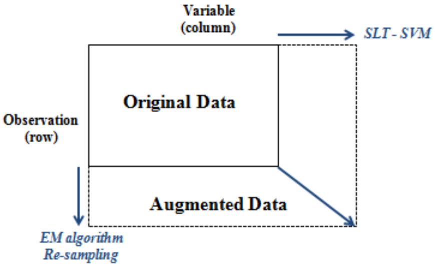

Fig. 2. Data augmentation by SVM and EM algorithm

First, DA augments the number of observations using expectation-maximization (EM) algorithm or re-sampling technique [4]. This is a row leveled DA approach. Second, DA transforms original data to augmented data with higher dimension. This is a column-leveled DA approach. DA is another approach for efficient data analysis. One of the column leveled DA approaches for classification is SVM [7],[9]. SVM is the classification version of SLT [16]. The groundwork of SLT was Vapnik-Chervonenkis (VC) theory [1]. SLT gives global optimization by convex searching and empirical risk minimization (ERM) [7]. In SLT, the original data are mapped onto a very high dimensional space using a kernel function. This is a popular approach in various DA methods. In this paper, we use SVM as a DA approach for classification task. SVM finds a hyperplane which is an optimal decision boundary using support vectors and margins. Figure 3 shows a maximum margin hyperplane in SVM.

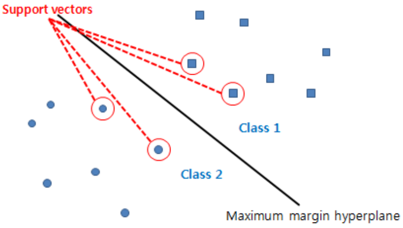

Fig. 3. Maximum margin hyperplane

SVM uses support vectors for constructing classifiers. In the case of s nonlinear class boundary of SVM, a slack variable is used for finding the optimal hyperplane. We solve the optimization problem as follows [9]:

$$t_i(wx_i+b)\geq 1-\xi_i\ ,\ \xi_i\geq 0\ ,\ i=1,2,...,n$$
$$Minimize\ \ J(w,\xi)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i \tag{5}$$

Where ξ is a slack variable. The *t* and *b* represent class and bias respectively. Also, *C* and *w* are a regularized constant and a weight vector. We use a kernel function for mapping from the original data space to the feature space with higher dimension. We use a radial basis function (RBF) as a kernel function in this paper [8]:

$$K(x,y)=\exp\left(\frac{-(x-y)^2}{2\sigma^2}\right), \tag{6}$$

where σ^2 is a kernel parameter of RBF which is a variance of Gaussian distribution. We knew that the PCA had the loss of information of data in the process of DR. Also, the SVM used only the support vectors of data for performing classification. Therefore, we need another approach to efficient classification. Next, we propose a classification method using DR.

3. Classification Method using Data Reduction

Classification is to assign a pattern (object) to one of the classes when a pattern is detected [17]. Also, classification contains the process of constructing a predictive model from given data to find the optimal class of new pattern. In this paper, we propose a method for efficient classification. This is combined Gaussian mixture model [13] and *K-NN*. In the process using the Gaussian mixture model, the dimension of input vector *X* is reduced by the number of categories of target variable *Y*. We divided the data into non-overlapping slices by the finite number of categories of *Y*. By performing the finite mixture model of Gaussian density, the distribution of input vector is approximated to any slice for obtaining kernel matrix of corresponding component means as following:

$$f(X\mid Y)=\sum\phi P(\mu,\Sigma), \tag{7}$$

where ϕ is the weight and *P(·)* is the Gaussian density with mean μ and covariance matrix *Σ*. The dimension of original data is reduced to the number of slices by the finite Gaussian mixture. Also, we construct optimal Gaussian mixture by Bayesian Information Criterion (BIC) [10]:

$$BIC=-2\log L+d\log N, \tag{8}$$

where *L* is likelihood function, *d* is the number of free estimated parameters, and *N* is a data size. We select the optimal Gaussian mixture model with the largest BIC. This research is comprised of the following steps:

X: input vector, *Y*: target variable

S1. Select *S* non-overlapping slices (finite number of categories of *Y*).
S2. Construct Gaussian mixture model using EM algorithm and BIC measure.
S3. Obtain kernel matrix from the estimated means of each mixture in the slices.
S4. Build generalized eigen-decomposition of obtained kernel matrix by covariance matrix of *X*.
S5. Estimate the subspace by the corresponding eigen-vectors.
S6. Reduce the dimension from original data space to subspace.
S7. Search the *K* nearest neighbors from *X*.
S8. Classify *X* by the most frequency class of the collection (*K* nearest neighbors)

So, this research develop a hybrid model combined Gaussian mixture model and *K-NN* algorithm for efficient classification. In a previous data mining, Gaussian mixture model was used for clustering, however we apply the mixture model to classification task.

4. Experiments and Results

In this paper, we perform the classification task for comparing the performances of the new DR, original dimension, DR and DA. Figure 4 shows the process of our comparison studies.

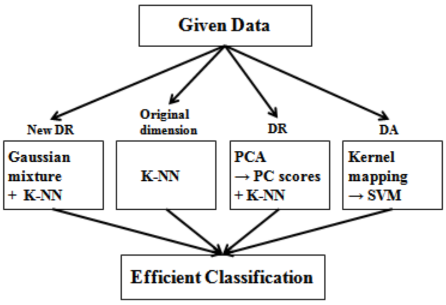

Fig. 4. Process of comparison study

We performed four separate analyses on each given data set. First, we showed the experimental result by the proposed methods. Second, we made experiment using original dimension data. We applied *K-NN* to the data with original dimension and get the results of accuracy (misclassification rate). Third, we constructed the PC score data from the PCA result after DR. We performed classification using the PC score data with low dimension. Fourth, we constructed a classification model of DA using SVM and measured the accuracy as criteria of the evaluated results of new DR, original dimension, DR and DA. To compare the performances of the comparative approaches, we analyzed data sets from the UCI machine learning repository [14]. Table 1 shows the summary of the searched databases for our analyses.

Table 1. Summary of data sets

Database	# of objects	# of variables	# of classes
Iris	150	4	3
Glass identification	214	9	6
Breast Cancer	569	32	2
Letter recognition	20000	16	26

We used two-thirds of given data as a training data set for constructing classification model, and the remaining data was used as a test data set for validation of the classification result [18]. We used R language in this experiment [19]. First, we got the compared results using popular Iris data set. Table 2 shows the accuracy and computing time. The number of PC for DR PCA was two. We determined the number having over 90% variance, also the number of support vectors (s.v.) of DA SVM was 55. It is seen that the misclassification rates of DR and DA were smaller than new DR and original dimension. However, the rates were not significantly different. The computing times were similar in all dimensions because the data size of Iris was small.

Table 2. Classification result: Iris data

Verifying performance	Misclassification rate (%)	Computing time (seconds)
New DR	6.00	0.20
Original dimension	6.00	0.34
DR (# of PC = 2)	4.00	0.22
DA (# of s.v. = 55)	4.00	0.07

Next, we analyzed the Glass identification data set. It has a larger sample size than the Iris data set. Table 3 shows that the misclassification rate of DR was the largest, because DR had loss of population information in the process of PCA. We found the accuracy of new DR was the best. Furthermore, the computing time of new DR was faster than original dimension and DR. The reason why the computing time of DA was the fastest is that SVM used only selected support vectors (small data) for performing classification.

Table 3. Classification result: Glass identification data

Verifying performance	Misclassification rate (%)	Computing time (seconds)
New DR	23.94	0.27
Original dimension	32.39	0.31
DR (# of PC = 4)	36.62	0.45
DA (# of s.v. = 127)	26.76	0.14

In Table 4, the Breast cancer Wisconsin (diagnostic) data set was analyzed. The number of objects and variables were larger than Iris and Glass identification data.

Table 4. Classification result: Breast cancer Wisconsin data

Verifying performance	Misclassification rate (%)	Computing time (seconds)
New DR	0.98	0.33
Original dimension	1.58	1.15
DR (# of PC = 2)	7.98	0.42
DA (# of s.v. = 101)	1.05	0.16

Similar to the results of the Glass identification data, the performance of DR was bad, and new DR had the best in accuracy. Also, the computing time of new DR was faster than original dimension and DR.

Table 5. Classification result: Letter recognition data

Verifying performance	Misclassification rate (%)	Computing time (seconds)
New DR	4.41	119.38
Original dimension	4.71	143.68
DR (# of PC = 11)	5.85	121.58
DA (# of s.v. = 7517)	5.72	33.85

Our last analyses by Letter recognition data set gave us the confirmation of the performance of new DR in Table 5. Since DA took the fastest computing time, it could be seen that DA approach has the best performance with respect to computing time. However, it was because DA by SVM did not use total data. In SVM classification, only a few support vectors were used as actual data. So, the computing time of DA classification was faster than the others. The accuracy results of DA were also good. The misclassification rate of new DR was better than other methods. We could verify the improved performance of our research.

Table 6. Best method for each data set

Data set	Accuracy performance
Iris	PCA, SVM
Glass identification	New
Breast cancer Wisconsin	New
Letter recognition	New

Table 6 shows which of the comparative methods is best in our classification. We conclude the proposed method provided the best performance in three data sets. Therefore, this research contributed an efficient approach to the classification task.

5. Conclusions and Future Works

In this paper, we proposed a new efficient DR method combining Gaussian mixture model and *K-NN* algorithm to construct a classification method. Our new DR method was compared with traditional DR and DA methods using PCA and SVM because our model was based on the traditional DR. The misclassification rate and the computing time were used for comparing the performance in the task of classification. PCA (DR approach) was considered as a good solution to avoid the curse of dimensionality, and SVM (DA approach) was also novel method for classification because this searched the optimal solution by convex optimization based on VC-dimension. Also, our method was compared to the original dimension of given data. From the most of experimental results, we found that the new DR approach was better than the original dimension, DR, and DA in accuracy (misclassification rate).

In our future work, we would develop diverse methods between DR and DA for classification, regression, and clustering. Also, we could apply diverse dimension approaches such as dimension augmentation vector machine and import vector machine to learning problems for classification, regression, and clustering.

References

[1] J. Han, M. Kamber, *Data Mining Concepts and Techniques*, 2nd edition, Morgan Kaufmann, 2006.

[2] R.A. Johnson, D. W. Wichern, *Applied Multivariate Statistical Analysis*, 3rd edition, Prentice Hall, 1992.

[3] J. H. Friedman, “On Bias, Variance, 0/1-loss, and the Curse of Dimensionality,” *Data Mining and Knowledge Discovery* vol. 1, pp. 55–77, 1997.

[4] M. A. Tanner, Tools for Statistical Inference, Springer, 1996.

[5] Y. Youk, S. Kim, Y. Joo, “Intelligent Data Reduction Algorithm for Sensor Network based Fault Diagnostic System,” International Journal of Fuzzy Logic and Intelligent Systems, vol. 9, no. 4, pp. 301-308, 2009.

[6] J. Keum, H. Lee, M. Hagiwara, “A Novel Speech/Music Discrimination Using Feature Dimensionality Reduction,” *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 10, no. 1, pp. 7-11, 2010.

[7] V. Cherkassky, F. Mulier, *Learning from data, Concepts, Theory, and Methods*, John Wiley & Sons, 1998.

[8] I. Oh, *Pattern Recognition*, Kyobo, 2008.

[9] V. N. Vapnik, *Statistical Learning Theory*, Wiley, 1998.

[10] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning, data mining, inference, and prediction*, Springer, 2001.

[11] N. G Polson, S. L. Scotty, “Data Augmentation for Support Vector Machines,” *Bayesian Analysis*, vol. 6, no. 1, pp. 1-24, 2011.

[12] L. Scrucca, “Model-based SIR for dimension reduction,” *Computational Statistics and Data Analysis*, vol. 55, pp. 3010-3026, 2011.

[13] K. C. Li, “Sliced inverse regression for dimension reduction,” *Journal of the American Statistical Association*, vol. 86, pp. 316-342, 1991.

[14] UCI ML Repository, http://archive.ics.uci.edu/ml/

[15] P. Giudici, *Applied Data Mining, Statistical Methods for Business and Industry*, Wiley, 2003.

[16] V. Cherkassky, F. Mulier, *Learning from data Concepts, Theory, and Methods*, John Wiley & Sons, 1998.

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

[18] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[19] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, http://www.R-project.org, 2001.

Daiho Uhm
He received his Ph.D. in department of statistics, Florida State University, U.S.A. in 2007, and BS and MS degrees in department of statistics, Inha University in 1997 and 1999, respectively. Currently he is a visiting assistant professor in department of statistics, Oklahoma State University. He is interested in survival analysis and computational statistics.

Phone : +1-405-744-5684
Fax : +1-405-744-3533
E-mail : daiho.uhm@okstate.edu, daiho.uhm@hotmail.com

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 in 2007. He is currently Associate Professor in department of Statistics, Cheongju University, Korea. He has researched data mining and management of technology (MOT).

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

Seung-Joo Lee
He received the BS degree in department of applied statistics from Cheongju University, Korea in 1985. Also, he received MS, and PhD degrees in department of Statistics, Dongkuk University, Korea, in 1987 and 1995. He is currently Professor in department of Statistics, Cheongju University, Korea. He has researched Bayesian statistics and multi-variate analysis.

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