International Journal of Fuzzy Logic and Intelligent Systems, vol. 9, no. 2, June 2009 pp. 141-146

Support Vector Machine based on Stratified Sampling

Sunghae Jun

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

Abstract

Support vector machine is a classification algorithm based on statistical learning theory. It has shown many results with good performances in the data mining fields. But there are some problems in the algorithm. One of the problems is its heavy computing cost. So we have been difficult to use the support vector machine in the dynamic and online systems. To overcome this problem we propose to use stratified sampling of statistical sampling theory. The usage of stratified sampling supports to reduce the size of training data. In our paper, though the size of data is small, the performance accuracy is maintained. We verify our improved performance by experimental results using data sets from UCI machine learning repository.

Key Words : Support Vector Machine, Stratified Sampling, Classification, Computing Cost

1. Introduction

Generally there are three learning approaches in data mining[1],[2]. They are classification, regression, and clustering. One of them, classification is a learning method to predict class labels[3]. Many algorithms have been proposed and used in the classification works[4],[5],[6]. Recently the most popular algorithm for classification is support vector machine(SVM)[7],[8],[9],[10]. SVM is a classification algorithm based on statistical learning theory(SLT) such as Vapnik-Chervonenkis(VC) dimension[11],[12],[13]. The algorithm has supported good results in the data mining[14],[15],[16]. But SVM has some problems[17],[18],[19]. A problem is its computing time. It causes difficulty to use SVM in the dynamic and online systems. To settle the problem, in this paper, we propose the usage of stratified sampling in the SVM modeling. That is, the stratified sampling reduces the size of training data for constructing SVM model. There is a work about SVM and stratified sampling[20]. This was a specific research studied on an imbalanced problem. But the goal of our paper is to reduce the computing time of SVM. This is general settlement of a SVM problem, its computing cost. Therefore we are able to improve SVM based on stratified sampling for decreasing its computing cost. SVM has had computing cost of training, though it has shown many good results in machine learning fields. For example, autonomous and intelligent agents need to get learning results immediately[14],[15]. But the computing time may be increased in proportion to the size of training data. So we use stratified sampling technique to reduce the training time in our research. Some studies about reducing computing time of SVM have researched[7],[14]. They were mainly bootstrap or dimension reduction[17]. But our work is different than these previous researches. An improvement of our study is to use stratified sampling for decrease of computing time of SVM. The contribution of our research is a simple and easy approach to solve the computing cost of SVM using a statistical sampling theory, stratified sampling. All classification works by SVM are based on SLT. SLT was introduced by V. Vapnik to overcome local optima and over-fitting problems of machine learning algorithms[13]. It is a theoretical algorithm of learning from given data. Recently new types of SLT based learning algorithms which are SVM, support vector regression(SVR), and support vector clustering(SVC) have been proposed. They are able to make tools for the theoretical analysis and practical algorithms. SVM, SVR, and SVC are good analytical methods for classification, regression, and clustering respectively. SLT describes statistical estimation with small samples effectively. Also, this theory includes classical statistical methods which are developed for large samples and strict parametric assumptions. SLT is based on empirical risk minimization(ERM) principle and VC dimension. The empirical risk is the average risk for the training data. This is minimized by choosing the appropriate parameters. In the supervised learning, the risk functional R(w) is minimized over the class of approximating function F(x,w). Where, the x and w are input and weight vectors respectively. The optimal parameter values are found by minimizing the empirical risk with respect to *w*. The theory of convergence of $R_{emp}(w)$ to R(w) includes bounds on the rate of convergence, which are based on VC dimension[12]. The VC dimension is a scalar value that measures the capacity or expressive power of a set of functions realized by the learning machine. VC theory explicitly takes into account the sample size and provides quantitative description of the trade-off between the model complexity and the available information. The VC dimension of a set of functions is *p* if and only if there exists a set of points such that these points can be separated in all 2^p possible placements[9]. This paper makes an attempt to use stratified sampling methods for solving the computing cost of various machine learning algorithms. As the first step of the attempts, we use the stratified sampling methods to reduce the computing time of SVM. To verify improved performance of our proposed research, we compare the computing times of stratified

Manuscript received Dec. 1. 2008; revised Jun. 8. 2009.

sampling data and total training data using four data sets from UCI machine learning repository[21]. Of course, we also draw a comparison the accuracies between the stratified sampling data and total data without any sampling. In our paper, though the size of data is small, we find that the performance accuracy is maintained like training total data. In the next section, we describe the SVM for classification and stratified sampling. Our proposed method is shown in the section 3. The section 4 shows the experimental results to verify our improved performances. Our conclusion and future work are presented in the final section.

2. SVM and stratified sampling

2.1 Support vector machine and Classification

SVM is a classification algorithm based on SLT. This is an analytical tool for efficiently training for the linear learning machines in the kernel-induced feature spaces, while respecting the insights provided by the generalization theory and exploiting the optimization theory[10]. A simple SVM is the maximal margin classifier which works only for linearly separable data in the feature space. The maximal margin classifier of SVM is to find the maximal margin hyperplane in an approximately chosen kernel-induced feature space. This is a convex optimization problem which is to minimize a quadratic function under linear inequality constraints. Also suppose we have some hyperplane which separates the positive from the negative examples. The points x which lie on the hyperplane satisfy wx+b=0, where w and b are weight and bias respectively. b/|w| is the perpendicular distance from the hyperplane to the origin, and |w| is the Euclidean norm of w. For the linearly separable case, SVM looks for the separating hyperplane with maximal margin. Now consider the points for which the equality of hyperplane hold requiring that there exists such a point is equivalent to choosing a scale for w and b. The points lie on the hyperplane with normal w and perpendicular distance from the origin. The solution of SVM modeling is performed by Lagrangian formulation. Because of the constraints will be replaced by constraints on the Lagrange multipliers. This is a crucial property which will allow us to generalize the procedure to the nonlinear case. We can introduce positive Lagrange multipliers , one for each of the inequality constraints. The rule is that for constraints of the form. The constraint equations are multiplied by positive Lagrange multipliers and subtracted from the objective function, to form the Lagrangian. For equality constraints, the Lagrange multipliers are unconstrained. Therefore SVM is a learning method based on SLT which has VC dimension and structural risk minimization(SRM)[12],[13]. This has been used as a popular classification tool. Also the tool is an analytical method for optimal training. In SVM, x is input vector and y has target labels which are +1 or -1. We predict the label of y according to x using the following[4].

$$y = sign(w \bullet x + b) \tag{1}$$

Where, w and b are weights and bias respectively. So $w \bullet x + b$ can predict y by a decision function sign(). We model SVM in the following formula.

$$\begin{aligned} & y_i (w \bullet x_i + b) \geq 1 \\ & \min_{w,b} \|w\| \end{aligned} \tag{2}$$

SVM classifier is to determine the maximal margin hyperplane in feature space from data space induced by kernel function. Support vectors are the points of training data to determine the labels of y. The support vectors are X_i in the following formula.

$$\begin{aligned} & y_i (w \bullet X_i + b) = 1 \\ & w = \sum \alpha_i y_i X_i \end{aligned} \tag{3}$$

α_i be Lagrange multiplier. Therefore, we are able to predict y about x by the following formula.

$$y = sign\left(\sum \alpha_i y_i X_i x + b \right) \tag{4}$$

X_i and x are support vectors and input vector respectively. Also this is equal to the following function.

$$f(x) = sign\left(\langle w, \phi(x) \rangle + b \right) \tag{5}$$

$\phi(x)$ and < , > are mapping from data space to feature space and dot product in the feature space. In the SVM, the dimension of feature space is extremely higher than input data space. In the above formula, the optimal values of w and b are computed by the next optimization equation using SRM.

$$\begin{aligned} & min. \qquad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \\ & s.t. \quad y_i \left(\langle w, \phi(x) \rangle + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{6}$$

C and ξ_i are regularization constant and slack variable respectively. This problem is solved by dual form.

$$\begin{aligned} & min. \quad L = -\sum_{i=1}^{N} \alpha_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \\ & s.t. \qquad \sum_{i=1}^{N} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C \end{aligned} \tag{7}$$

The k(,) is a kernel function which maps input vector into high-dimensional feature space.

$$k(x_i, x_j) = \left\langle \phi(x_i), \phi(x_j) \right\rangle \tag{8}$$

According to grid searching, evolutionary computing, and so on, the regularization constant and kernel parameters have been determined[16].

2.2 Stratified Sampling

The goal of sampling is to maintain the information of the population for a given cost[22],[23]. Simple random sampling(SRS) has shown good performance in the general population not depended on any classes of given training data[24]. But there are some label characteristics in the

population, the performances are decreased[25]. Stratified sampling is a gate to solve this problem of SRS. Stratified sampling is a sampling method that separates the population elements into non-overlapping strata, and then performing SRS from each stratum[23]. The following shows how to draw a stratified sample[26],[27].

(Step1) Specify the strata

Each sampling unit is assigned into a proper stratum.

(Step2) Divide sampling units into strata

Simple random sampling from each stratum is performed after the sampling units are divided into strata.

Also, the samples from the strata are independent. The terms about stratified sampling are presented in the followings.

$$N = N_1 + N_2 + \ldots + N_L \quad (9)$$

Where, L, N, and N are the numbers of strata, sampling units in stratum I, and sampling units in the populations.

Stratified sampling is that the population is partitioned some strata which are not overlapping, then simple random samplings perform each stratum[22]. In stratified sampling, the smaller the data size is, the smaller the variance is. The stratified sampling method has been used in many computing system for solving diverse problems of them[24],[25],[26],[27]. In this paper we do 10% sampling equally, because the issue of our paper is not variance estimation but computing time.

3. SVM based on Stratified Sampling

To reduce the computing time of SVM, we consider stratified sampling. The data size may be decreased by efficient sampling approach. Simple random sampling is not suitable in classification data. That is because the data points with sparse labels are not able to be sampled from total training data set. So, in this paper, we sample according to each label in total training data set. This is stratified sampling. For learning SVM, we partition total training data into k disjoint classes, (C_1, C_2, …, C_k). These are called strata. If the number of points in total training data is N, we define the size of each stratum as the following.

$$n_1 + n_2 + \cdots + n_k = N \quad (10)$$

So the class and its data size are shown in the following pairs, (class, size of class).

$$(C_1, n_1), (C_2, n_2), \cdots, (C_k, n_k) \quad (11)$$

We perform 10% simple random sampling from each stratum. Therefore we get the 10% stratified sample from the total training data set as the following.

$$(C_1, rn_1), (C_2, rn_2), \cdots, (C_k, rn_k) \quad (12)$$

Where, rn_i is the size of *i*th stratum after stratified sampling. The following figure shows our work briefly.

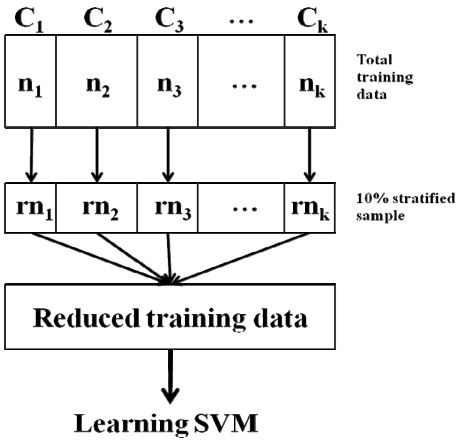

Fig. 1 SVM learning by stratified sampling.

Where c_i and n_i are *i*th class label and stratum size of the training data. Above rn_i is equal to $0.1*n_i$. That is, rn_i is 10% simple random sampling from C_i with data size n_i. So our reducing process of SVM computing time is performed as the following steps.

T : total training data
N : size of the total data
C_i : *i*th label(class) of the data, or *i*th stratum
n_i : size of C_i
rn_i : size of 10% stratified sampling data from class C_i with data size n_i

(Step1) Define strata (in total data)

The number of labels(classes) is the size of strata.

$T=(C_1, C_2, \ldots, C_k)$, k: the number of labels(classes)
$N=n_1+n_2+\ldots+n_k$

(Step2) Sample each stratum (in each class of total data)

10% simple random sampling is perform in each stratum.

$rn_i=0.1*n_i$
$0.1*N=0.1*(n_1+n_2+\ldots+n_k)$
(Step3) Model SVM (in 10% stratified sample)

To reduce the computing time, SVM is modeled in 10% stratified random sample. In this paper, we use 10% stratified sampling heuristically to compare the computing times of two data sets. The experimental results according to different percentages of stratified sampling will be studied in the future. A contribution of our work is to use statistical sampling method(stratified sampling) for solving the computing cost of machine learning algorithm(SVM). To apply more diverse sampling approaches into some problems of machine learning algorithms are our future works.

4. Experimental Results

To verify our improved performances, we make experiments using data sets from UCI machine learning repository[21]. The

data sets consist of two types which have large and small numbers of classes. The criterion between large and small is the size of classes of the output variables. In this paper, we determine the size as 10 heuristically. So Adult and Iris plants data sets are considered as small size data. Similarly the large size data sets are Letter image recognition and Protein location site data. We can consider the other size of classes. Also, the experimental results are able to be found by the different sizes. These will be our future works. The following table shows the information of the data sets for our experiments.

Table 1. Data information

Data set	# of classes	# of points
Adult	2	32561
Iris plants	3	150
Letter image recognition	26	13333
Protein location site	10	1484

The task of Adult data set is to predict whether income exceeds $50,000 per year(2 classes) by census input variables which are age, education, capital gain and loss, and so forth. Iris data set is the most popular database in the classification fields. The data set contains 3 classes of 50 instances each type of iris plant. The goal of Letter image recognition is to classify each of many black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The data set of Protein location site is to predict the localization site of protein. In the above, the numbers of classes of adult and iris plants data sets have 2 and 3 respectively. So we set them as the small class types. Likewise, the letter image recognition and protein location site data sets have large class types. In our experiments, we find the relation between the number of classes and the accuracy in stratified sampling data. Of course, we compare the computing times between stratified sampling used and not used data sets. Our experiments are performed on a Intel duo CPU(2.66GHz) with 2GB RAM memory of the Window XP system. Also, to perform the SVM and stratified sampling, we use e1071 and sampling packages in R-project[28]. We show the experimental result in the following tables.

Table 2. Regularization constant and kernel parameter

Data set	Regularization Constant	Kernel Parameter
Adult	1	0.16670
Iris plants	1	0.25000
Letter image recognition	1	0.06250
Protein location site	1	0.00068

Above table II shows the values of the regularization constant and the parameter of kernel function of SVM. We use default values of robust grid search[20] though there are some methods of optimal selections for the values[16], because the goal of this paper is to compare computing times between stratified sampling used and not used in equal condition. We find the experimental results which are the numbers of support vectors(s.v.), accuracies, and computing times between total training data and 10% stratified sampling data. To decrease the computing time soundly, we considered 10% stratified sampling. But, the lower percentile stratified sampling than 10% brought out not good accuracy. So in this paper, we used the 10% stratified random sampling.

Table 3. Experimental results

Data set	training	# of s.v.	Accuracy (%)	Computing time (seconds)
Adult	Total	12517	82.4	191.28
	Stratified	1372	82.3	1.77
Iris plants	Total	41	96.1	0.07
	Stratified	28	96.1	0.05
Letter image recognition	Total	7422	94.3	29.78
	Stratified	1136	79.8	1.30
Protein location site	Total	1481	36.3	2.27
	Stratified	136	30.4	0.37

In table 3, the 'total' is to use total data set without any sampling. And the 'stratified' represents the usage of 10% stratified sample to construct SVM model. The results of accuracy were gained by 10 folds validation. Also, # of s.v. is the number of support vectors. The result of Adult data set shows the extreme reduction of computing time from total training data set(191.28 seconds) to 10% stratified sampling data set(1.77 seconds). In addition the accuracies between total data and 10% stratified sampling data are similar. So we get good result from the experiment of Adult data set. In Iris plants data, the decrease of computing time of stratified sampling was small. The accuracy performances of total and stratified sampling data sets are same(96.1%). Also we get the decrease of computing time in Letter image recognition. But the accuracy of stratified sampling is poor. By this result, we know the accuracy performance is depended on the number of classes. The experimental result of Protein location site data is similar to the result of letter image recognition data. By our example studies, we know improved performance of SVM using stratified sampling when the number of points is large and the number of classes is small. Therefore, in the large data sets with small classes, we propose to use stratified sampling for reducing the computing time of SVM in classification.

5. Conclusions and future works

In this paper, we proposed the usage of stratified sampling to overcome the computing cost of SVM model. Though SVM is

good learning algorithm, it has computing cost problem in classification works of large data. Some works have studied to reduce the computing time of SVM. We considered a new approach for solving the computing cost of SVM in our research. Using 10% stratified sampling, the data set for constructing SVM were reduced soundly maintaining the performance of accuracy. So we showed computing time reduction of SVM using stratified sampling. In future works, we will make experiments using more data sets by simulation. The sampling size will be also considered in our future study. Finally, diverse sampling techniques including stratified sampling will be applied to many machine learning algorithms to reduce their computing times.

References

[1] P. Ciudici, *Applied Data Mining*, Wiley, 2003.

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

[3] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data mining, Inference, and Prediction, Springer, 2001.

[4] S.R. Gunn, "Support Vector Machines for Classification and Regression", *Technical Report*, University of Southampton, 1998.

[5] V. Cherkassky, F. Mulier, *Learning From Data Concepts, Theory, and Methods*, John Wiley & Sons, 1998.

[6] J. H. Friedman, "An Overview of Predictive Learning and Function Approximation," *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, vol. 136, Springer, 1994.

[7] Y.S. Jia, C.Y. Jia, and H.W. Qi, "A New Nu-Support Vector Machine for Training Sets with Duplicate Samples," *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, pp. 4370-4373, 2005.

[8] C. Gold, P. Sollich, "Model selection for support vector machine classification", *Neurocomputing*, vol. 55(1-2), pp. 221-249, 2003.

[9] S. Haykin, *Neural Networks*, Prentice Hall, 1999.

[10] C. Nello, S.-H. John, An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000.

[11] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[12] V.N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.

[13] V.N. Vapnik, "An Overview of Statistical Learning Theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988-999, 1999.

[14] J. Wang, X. Wu, and C. Zhang, "Support vector machines based on K-means clustering for real-time business intelligent systems," *Int. J. Business Intelligence and Data Mining*, vol. 1, no. 1, pp. 54-64, 2005.

[15] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V.N. Vapnik, "Support Vector Clustering", *Journal of Machine Learning Research*, vol. 2, pp. 125-137, 2001.

[16] S.-H. Jun, "Web Usage Mining Using Evolutionary Support Vector Machine", *Lecture Note in Artificial Intelligence*(LNAI, AI'2005), vol. 3809, pp. 1015-1020, Springer-Verlag, 2005.

[17] L. Xuchun, Z. Yan, and E. Sung, "Sequential bootstrapped support vector machines-a SVM accelerator," *Proceedings of IEEE International Joint Conference on Neural Networks*, vol. 3, pp. 1437-1442, 2005.

[18] F. Friedrichs, C. Igel, "Evolutionary Tuning of Multiple SVM Parameters", *Proceedings of the 12th European Symposium on Artificial Neural Networks*, 2004.

[19] P. Ling, Y. Wang, N. Lu, J. Y. Wang, S. Liang, C. G. Zhou, "Two-Phase Support Vector Clustering for Multi-Relational Data Mining", *Proceedings of the International Conference on Cyberworlds*, 2005.

[20] Sd F. Vilarino, P. Spyridonos, J. Vitria, P. Radeva, "Experiments with SVM and Stratified Sampling with an Imbalanced Problem: Detection of Intestinal Contractions," *LNCS*, vol. 3687, pp. 783-792, 2005

[21] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/

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

[23] R. L. Scheaffer, W. Mendenhall III, R. Lyman Ott, *Elementary Survey Sampling*, Fifth Edition, Duxbury Press, 1996.

[24] 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.

[25] 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:Applications and Reviews*, vol. 37, no. 6, pp. 1223-1230, 2007.

[26] 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.

[27] 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.

[28] The R Project for Statistical Computing, http://www.r-project.org

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