The Fourteenth International Symposium on Artificial Life and Robotics 2009 (AROB 14th '09),
B-Con Plaza, Beppu, Oita, Japan, February 5 - 7, 2009

An analysis of epression data using Support Vector Machine and feature selection methods

Michifumi Yoshioka, Norihiro Shimoda, Sigeru Omatu
Dept. of conputer science and intelligent System
Osaka Prefecture University
Sakai, Osaka, Japan

Abstract

Gene expression data is signi cantly expected as aid in the development of efficient cancer diagnosis and classi cation platforms. However, Gene expression data is high dimensional and the number of samples is small in comparison to the dimension of the data, and furthermore, noisy inherently. Therefore, we would be better o using not all genes but part of genes in order to classify gene expression data. Previous works introduce a method of hybrid Genetic Algorithm and Support Vector Machine (GASVM) for nding the small subset of informative genes which maximize the classi cation accuracy. However, these methods have some problems. Firstly, previous works cost a large amount of calculation because of SVM using high dimensional data and its iteration. Secondly, previous works use only accuracy of SVM as a criterion for gene selection. This might cause over t-ting. Thus, we introduce a criterion named "Con - dence Margin". The criterion means a "goodness" of hyperplane made by SVM. Using " Con dence Margin ", proposed method cut o iteration, and, cost less amount of calculation than previous works. Furthermore, proposed method shows as a high accuracy as previous works.

1 introduction

DNA microarray has been the key technology in recently biology and helped us to understand the biological system because of its ability to monitor the expression levels of thousands of genes simultaneously. The expression levels mean the hereditary activities in each genes. Therefore, gene expression data is created by a process known as microarray and represents the activation level of each gene within an organism at a particular point of time, and the class label means which class a person belongs to. Thus, gene expression data is signi cantly expected as aid in the development of efficient cancer diagnosis and classi cation platforms.

This work deals with the way of classi cation using gene expression data. From the classi cation point of view, it is well known that, when the number of samples is much smaller than the number of features, classi cation methods may lead to data over tting, meaning that one can easily nd a decision function that correctly classi es the training data but this function may behave very poorly on the test data. In order to classify gene expression data, it is necessary to reduce the data dimension by selecting a subset of features that are relevant for classi cation, that is, we should not use all genes but part of genes which contribute to classi cation [1]. Previous works introduce a method of hybrid Genetic Algorithm and Support Vector Machine (GASVM) [2], and extended-GASVM (New-GASVM) [3]for gene selection and classi cation tasks. The methods consists of two parts. GA part works for feature selection and evaluation, that is, makes gene subsets and evaluate its goodness, and SVM part measures accuracy of gene subsets from GA part. Then, the accuracy is used to evaluate gene subsets in GA part. As seen above, the method using accuracy of a classi er as criterion is called " Wrapper tequniques "[4]. The method shows very high accuracy of the benchmark datasets. However, previous works have two problems. Firstly, previous works cost a large amount of calculation because of SVM for high dimensional data and LOOCV (Leave-One-Out Cross-Validation). Secondly, previous works use only accuracy of SVM as a criterion. Therefore, the result (optimal gene subset) show high accuracy for training data, but might not show high accuracy for unknown data, so-called " over tting ". In order to resolve these problems, we focus on data distribution in feature space, and select features so that two data constellation are isolated each other. In concrete terms, we use criterion named "Con dence Margin" [5]. The criterion means a "goodness" of hyperplane made by

SVM. Using "Con dence Margin", we have proposed a new method that selects genes and classi es gene expression data. Proposed method achieve drastically reduction of calculation amount because of cutting o iteration(LOOCV) when selected gene sbset are evaluated. Concequently, proposed method belongs to" Filter techniques "[4].

By the way, Filter tequniques carry out feature selection as preparation. Usually, lter techniques select features as preparation. As a result, lter techniques tend to show smaller amount of calculation than Wrapper techniques, but lower accuracy simultaneously. However, proposed method shows smaller amount of calculation and higher or equal accuracy than previous works which belongs to wrapper method.

2 Previous Works

2.1 GASVM

GASVM consists of two main components: GA and SVM classi er. The GA will select the subsets of features and then the SVM classi er evaluates the subsets during a classi cation process. The result of the classi cation process is used for the tness value of GA. We describe the outline of GASVM as below.

[individual representation] An individual represents a selecded gene subsets. The chromosome (individual) is represented by binary vector whose dimension are equal to the number of genes in gene expression data. If a bit is 1, it means that the corresponding feature is selected, and If a bit is 0, it indicates that the corresponding feature is not selected.
[making gene subsets] According to the rule described "individual representation", gene subsets are made of each chromosome by comparison with gene expression data.
[evaluation of tness] The tness function of each individual is determined by evaluating the SVM using a training set. Hence, this research is used with a tness function containing classi cation accuracy as mentioned below,

$$Fitness(x) = accuracySVM(x) \qquad (1)$$

where $accuracySVM(x)$ is the LOOCV accuracy of the classi er with the features subset selection which is represented by x.
[GA operation] Using tness, the individual is applied to some normal evolution steps in GA, that is, selection, crossover and mutation.

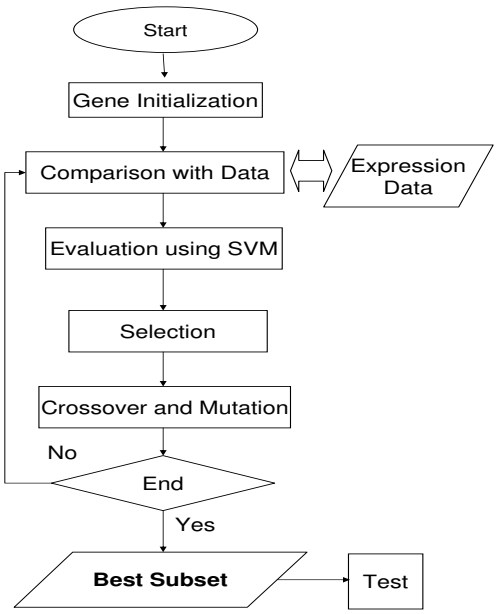

Figure 1: A ow chart of GASVM

The GA is used to maximize the tness value in order to nd the optimal features subset which has been achieved the highest LOOCV accuracy. The optimal subset from training set is used to construct SVM classi er.The Flow chart of GASVM are shown in Figure 1.

2.2 Difficulty of GASVM

GASVM investigate the optimal gene subset which maximizes the accuracy of SVM using the accuracy of SVM itself, and the number of samples is much smaller than the number of features in gene expression data in this case. Therefore, the selected gene subset lead to over tting, and lacks the foundation why these genes are optimal. Hence, it is a better way to select gene subset according to another criterion outside the accuracy of SVM ,and that result in high accuracy simultaneously.

3 Proposed Method

3.1 GASVM-CM

As an another criterion, we have employed the distance between two data constellation. There is some measurements which mean distance in SVM. Hence, taking into account the fact that gene expression data is linear inseparable and noisy, we adopt" Con dence Margin ". "Con dence Margin" is "Margin"[6] multiplied by "Con dence" [7]. Here, Con dence Margin is

$$ConfidenceMargin = Confidence \quad Margin \quad (2)$$

where $Margin$ means distance from hyperplane to support vector, that is, $Margin$ means the geometric distance between two data constellation. $Confidence$ is the distance which imposes penalties on the misclassi ed samples. Therefore, $ConfidenceMargin$ is the distance which takes into account the misclassi- ed sample, that is, it means goodness of hyperplane made by SVM. We have proposed a new method of gene selection and classi cation which use "Con dence Margin" in place of accuracy of SVM in GASVM, and named it GASVM-CM. The proposed method is used with a tness function as mentioned below.

$$Fitness(x) = ConfidenceMargin(x) \quad (3)$$

where $fitness(x)$ is tness value of individual x, and $ConfidenceMargin(x)$ is Con dence Margin of individual x. GASVM-CM remains basically the same with GASVM. However, it has the di erent way of evaluating individuals from GASVM. The outline of evaluating individuals in GASVM-CM are described as below.

[measurement of Margin] SVM are trained by gene subset made from each individuals. As a result, support vector and hyperplane are determined in each individuals, and then, Margin is able to measured.

[measurement of Con dence] We can measure distance from determined margin to each samples taking into account its class label, and then, average these distance. It is Con dence in each individuals.

[Calculation of Con dence Margin] Con dence Margin is calculated by Margin and Con dence from previous step.

The proposed method searches combination of genes (individual) which maximizes Con dence Margin.

4 Experiment

4.1 Benchmark datasets

The rst benchmark gene exression microarray dataset is Colon Cancer. The data contains expressin levels of 2000 genes from 40 tumor and 22 normal colon tissues. The dataset has only 62 samples for training data, originally analyzed by Alon et al[8] and downloaded from http://microarray.princeton.edu/oncology/ The second benchmark gene expression microarray dataset is Leukemia Cancer. The data contains examples of human acute leukemia, originally analyzed by Golub et al[9]. The dataset containing expression levels of 7129 genes can be obtained at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi . The bone marrow or blood samples were taken from 72 patients, 47 with acute myloid leukemia (AML) and 25 with acute lymphoblastic leukemia (ALL). The training data consists of 38 samples and the remaining 34 samples were used as testing data.

4.2 Coventional approaches

In order to evaluate performance of proposed method, we experiment on GASVM, GASVM-CM, and comventional approaches. We introduce two methods which use only Margin, and Con dence as a criterion respectively. Those are named GASVM-M (Margin only) and GASVM-C (Con dence only). The comventional approaches are used a tness function as mentoned below,

$$fitness(x) = Margin(x) \quad (4)$$

$$fitness(x) = Confidence(x) \quad (5)$$

where $fitness(x)$ a tness value of individual x. $Margin(x)$ is a Margin from individual x , $Confidence(x)$ is a Con dence from imdividual x.

4.3 Result of experiment and discussion

Tables 1. and 2. show the results of the experiments for Colon Cancer and Leukemia Cancer datasets. Number of genes means the number of optimal gene subset in each method. Accuracy means accuracy of SVM on test data using optimal gene subset from training data. However, only LOOCV procedure was used to measure the classi cation accuracy on Colon Cancer dataset because this data set had only the training set.

Table 1: The Experimental Result of Colon

	Number of genes	Run-time (minute)	Accuracy (%)
GASVM	11	1380	87.10
GASVM-CM	29	420	88.71
GASVM-M	17	8	72.58
GASVM-C	70	450	72.58

Table 2: The Experimental Result of Leukemia

	Number of genes	Run-time (minute)	Accuracy (%)
GASVM	5	600	82.35
GASVM-CM	33	240	88.24
GASVM-M	10	7	73.53
GASVM-C	93	240	70.59

Comparing GASVM with GASVM-CM rstly, GASVM-CM have achieved the higher accuracy than GASVM, and the less amount of calculation. This is the comparison between wrapper technique and Filter technique. E ectiveness of our proposed method is con rmed because it means that the proposed method is able to avoid over tting.

Comparing GASVM-CM with comventional approaches secandly, GASVM-CM have achieve the higher accuracy than comventional approaches, however, the greater amount of calculation basically. The result is also very contented because it means that the proposed method can higher accuracy than comventional approaches. The result about Run-time is appropriate because Con dence Margin are calculated by Margin and Con dence measured in advance. As mentioned above, e ectiveness of using Con dence Margin as a criterion of feature selection on SVM is con rmed.

5 Conclusion

This paper have introduced Con dence Margin and proposed a method using Con dence Margin. Proposed method belongs to Filter method. Generally, Filter techniques tend to show smaller amount of calculation than Wrapper techniques but lower accuracy simultaneously. However, proposed method shows smaller amount of calculation and higher or equal accuracy than previous works which belongs to wrapper method.

References

[1] M. S. Mohamad and S. Deris, "Feature selection method using genetic algorithm for the classi cation of small and high dimension data" *Proc. Int. Symp. Info. Com. Tech*, p.13-16(2004)

[2] Edmundo Bonilla Huerta, Beatrice Duval, and Jin-Kao, "A Hybrid GA/SVM Approach for Gene Selection and Classi cation of Microarray Data" Proc. ofEvoWorkshops 2006, LNCS 3907, p.34-44, 2006.

[3] MOHAD SAVERI MOHAD ad SAFAAI DERIS , "A HYBRID OF GENETIC ALGORITHM AND SUPPORT VECTOR MACHINE FOR FEATURES SELECTION AND CLASSIFICATION OF GENE EXPRESSION MICROARRAY" *International Journal of Computational Inteligence and Apprications*, Vol.5, No.1 pp.91-107, 2005.

[4] Yvan Saeys Inaki Inza, and Pedro Larranaga, "A review of feature selection techniques in bioinformatics", *Bioinformatics*, Vol.123, No.19 pp.91-177, 2007.

[5] K.Aoki, S. Kuroyanagi, M. KUGULER, A. S. NUGROHO, and A. IWATA, "Feature selection Using Con dent Margin for SVM", *IEICE TRANS D-II*, No.12 pp2291-2300, 2005.

[6] Nello Cistianini, John Shawe-Taylor, *An Introduction to Spport Vector Machines and other kernel-based learning methods*, Cambridge University Press, 2000.

[7] Ling Li, Amrit Pratap, Hsuan-Tien Lin, and Yaser S. abu-Mostafa, "Improving Generatio by Data Categorization", *PKDD 2005*, LNAI 3721, pp.157-168, 2005.

[8] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proc. Natnl. Acad. Sci. USA*, volume 96, 1999.

[9] T. R. Golub, D. K. Slnim, P. Tamayo, C. Huarg, M. Gaasenbeek, J. P Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloom- eld, and E. S. Lander, "Molecular classi cation of cancer: Class discovery and class prediction by gene expression monitoring", *Science*, 286:531-537, 1999.