Chapter 19
Microarray Data Analysis with Support Vector Machine

Si-Hao Du, Jin-Tsong Jeng, Shun-Feng Su, and Sheng-Chieh Chang

Abstract Microarray data analysis approach has became a widely used tool for disease detection. It uses tens of thousands of genes as input dimension that would be a huge computational problem for data analysis. In this chapter, the proposed approach deals with selection of feature genes and classification of microarray data under support vector machine (SVM) approach. Feature genes can be finding out according to the adjustable epsilon-support vector regression (epsilon-SVR) and then to select high ranked genes after all microarray data. Moreover, multi-class support vector classification (multi-class SVC) and cross-validation methods apply to acquire great prediction classification accuracy and less computing time.

Keywords Support vector machine • Support vector regression • Multi-class support vector classification • Feature genes • Microarray data analysis

19.1 Introduction

Machine learning is one of artificial intelligence, which has the ability to learn. Machine learning techniques have been successfully applied to cancer classification for microarray data [1]. In machine learning approach, one of popular approaches is support vector machine (SVM) that can deal with classification under support vector classification (SVC) and regression analysis under support vector regression (SVR) [2]. In recent years, "feature selection" became a popular topic. It means used some methods to find feature genes from original genes. In general, cost will

S.-H. Du • S.-F. Su
Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

J.-T. Jeng (✉)
Department of Computer Science and Information Engineering, National Formosa University, Yunlin County, Taiwan
e-mail: tsong@nfu.edu.tw

S.-C. Chang
Aeronautical Systems Research Division, National Chung-Shan Institute of Science and Technology, Taichung, Taiwan

J. Juang (ed.), *Proceedings of the 3rd International Conference on Intelligent Technologies and Engineering Systems (ICITES2014)*, Lecture Notes in Electrical Engineering 345, DOI 10.1007/978-3-319-17314-6_19

increase under the number of genes in disease detection. Besides, many studies focused on combined with feature selection and SVM to deal with that reduce gene number and classification [3–7]. Zhang et al. [3] proposed *r*-test methods convert gene ranking results into position *p*-values to evaluate the significance of genes. Tang et al. [4] purposed a new two-stage SVM-recursive feature elimination (SVM-RFE) algorithm what overcomes the instability problem of the SVM-RFE to achieve better algorithm utility. And then have demonstrated that the two-stage SVM-RFE is significantly more accurate and more reliable than the SVM-RFE. Kung and Mak [5] purposed a fusion strategy to integrate the diversified information embedded in the symmetric doubly supervised (SDS) formulation. However, simulation studies on protein sequence data for subcellular localization confirm that the prediction can be significantly improved by combining vector-index-adaptive SVM (VIA-SVM) with relevance scores (e.g., Signal-to-Noise Ratio (SNR)) and redundancy metrics (e.g., Euclidean distance). In Leung and Hung [6], a multiple-filter-multiple-wrapper (MFMW) approach is proposed that makes use of multiple filters and multiple wrappers to improve the accuracy and robustness of the classification, and to identify potential biomarker genes. Lee and Leu [7] purposed a novel hybrid method for feature selection in microarray data analysis. The method first uses a genetic algorithm with dynamic parameter setting (GADP) to generate a number of subsets of genes and to rank the genes according to their occurrence frequencies in the gene subsets. Second applies the χ^2-test for homogeneity to select a proper number of the top-ranked genes for data analysis. Finally, they use the classification of SVM to verify the efficiency of the selected genes. Based on the above description, there are many studies focused on the topic of feature selection, and get good experiment results for prediction. For the SVM, the most of results are used SVC to do classification. In this chapter, we apply SVM for the microarray data analysis, the process in feature selection with SVR and in classification with multi-class SVC. That is, the analysis of microarray, selection of feature gene, and classification all use SVM in this chapter.

19.2 Characteristic of Ovarian Microarray Data

There are 41 samples and each sample is a piece of the microarray. These microarray samples are divided into four classes; namely, normal ovaries class, benign ovarian tumors (OVT) class, ovarian cancers at stage I (OVCAI) class, and ovarian cancers at stage III (OVCAIII) class in the ovarian cancer microarray data. Tissues applied in this study included 6 normal ovaries class, 13 OVT class, 7 OVCAI class, and 15 OVCAIII class in Table 19.1. All ovarian cancer microarray procedures

Table 19.1 Category of ovarian microarray

	Normal	OVT	OVCAI	OVCAIII	Total
Sample number	6	13	7	15	41

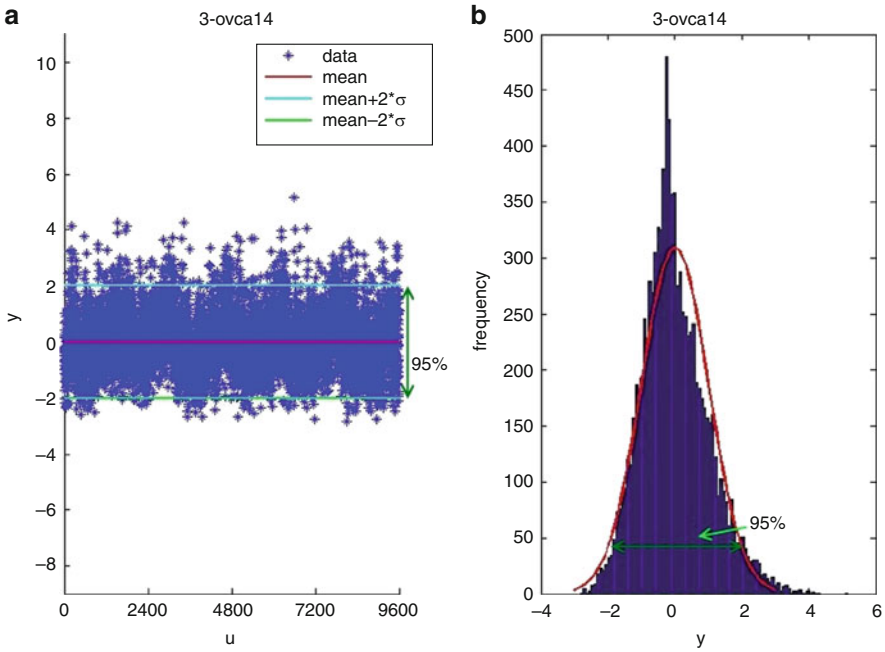

Fig. 19.1 Feature genes selection based on statistics theory

were performed in a dust/climate control laboratory at China Medical University. A sequence-verified human cDNA library containing 9,600 human cDNA clones was a kind gift from the National Health Research Institute of Taiwan [8].

Figure 19.1a shows the microarray data information where u is the number of genes in microarray and y is log(based 2) of R/G normalized ratio. R is magnitude of Cy5 and G is magnitude of Cy3. Traditionally, biologists found out feature genes based on statistics theory. The method is calculated p-value, based on mean and standard deviation. They usually use another 5 % genes to be feature genes for disease detection (see Fig. 19.1b). In general, the character of microarray data has a wave property from Fig. 19.1a. Therefore, the nature of microarray data is nonlinear. Hence, we proposed SVM that can deal with nonlinear problem to improve statistical method.

19.3 The Proposed Approach

SVM is a new classification and regression technique that was proposed by Vapnik, and successfully applied to many different fields [9]. The concept of SVM is that separate different high-dimensional labeled data according to optimal hyperplane. Besides, in SVM the data applies kernel function to map input data into another

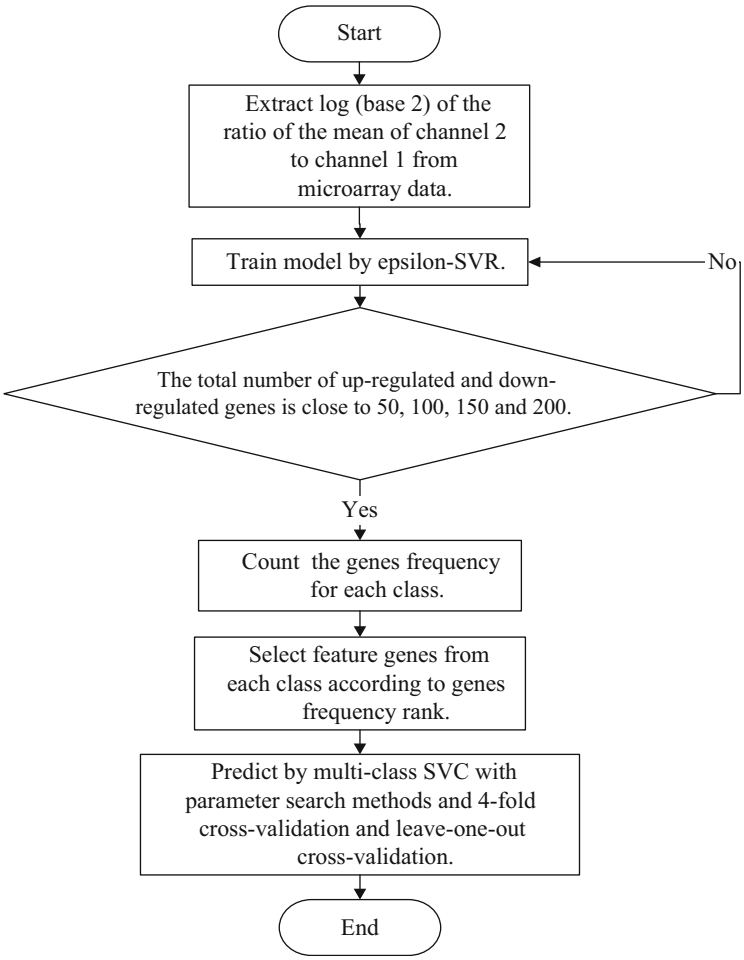

Fig. 19.2 The flowchart of the proposed approach

space. This chapter chooses radial basic function in multi-class SVC and epsilon-SVR. The flowchart of the purposed approach is shown in Fig. 19.2.

In general, the fluorescent dyes Cy3 (green) and Cy5 (red) are most often used to prepare labeled cDNA for microarray hybridizations. In this chapter, we only consider the magnitude of Cy5 and Cy3 in microarray data. Firstly, log(base 2) of the R/G ratio of the mean of channel 2 to channel 1 from microarray data is used. The genes expression data had been recorded in column named log 2 ratio normalized *R/G* mean as follows:

$$\text{Log(base2) of } R/G \text{ Normalized Ratio (Mean)} = \log_2 \frac{\text{Cy5}}{\text{Cy3}}. \tag{19.1}$$

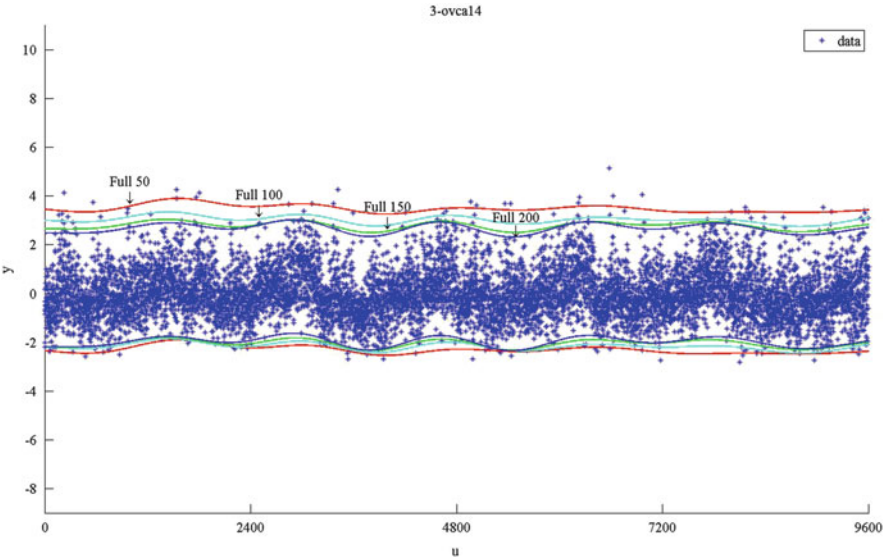

Fig. 19.3 The feature genes, finding out via the proposed SVR with different ε values

Secondly, based on adjust epsilon-SVR to build a smooth curve that can find out the total number of upregulated and downregulated genes is close to 50, 100, 150, and 200 that is shown in Fig. 19.3. The main concept of ε-SVR is proposed to find out the feature gene as in Fig. 19.4 under certain ε in SVR. If ε increased as red arrow then the total number of upregulated and downregulated would be reduced. The parameter ε could control how many genes would be filtered. Hence, using ε-SVR to filter out four classes of microarray data and record upregulated and downregulated genes is close to 50, 100, 150, and 200.

Thirdly, count and record the genes frequency for each class. For example: the gene named A, it was filtered five times in class 1, ten times in class 2, and three times in class 3 and recorded it into gene sets like "Full 50," "Full 100," "Full 150," and "Full 200". "Full 50" means a gene set that finds out the total number of upregulated and downregulated genes close to 50 with each sample from original genes. Fourth, select feature genes from each class according to genes frequency rank (from high to low, and if had existed then selected minor). Finally, use multi-class SVC with parameter search methods to classification microarray data according to fourfold cross-validation and leave-one-out (LOO) cross-validation.

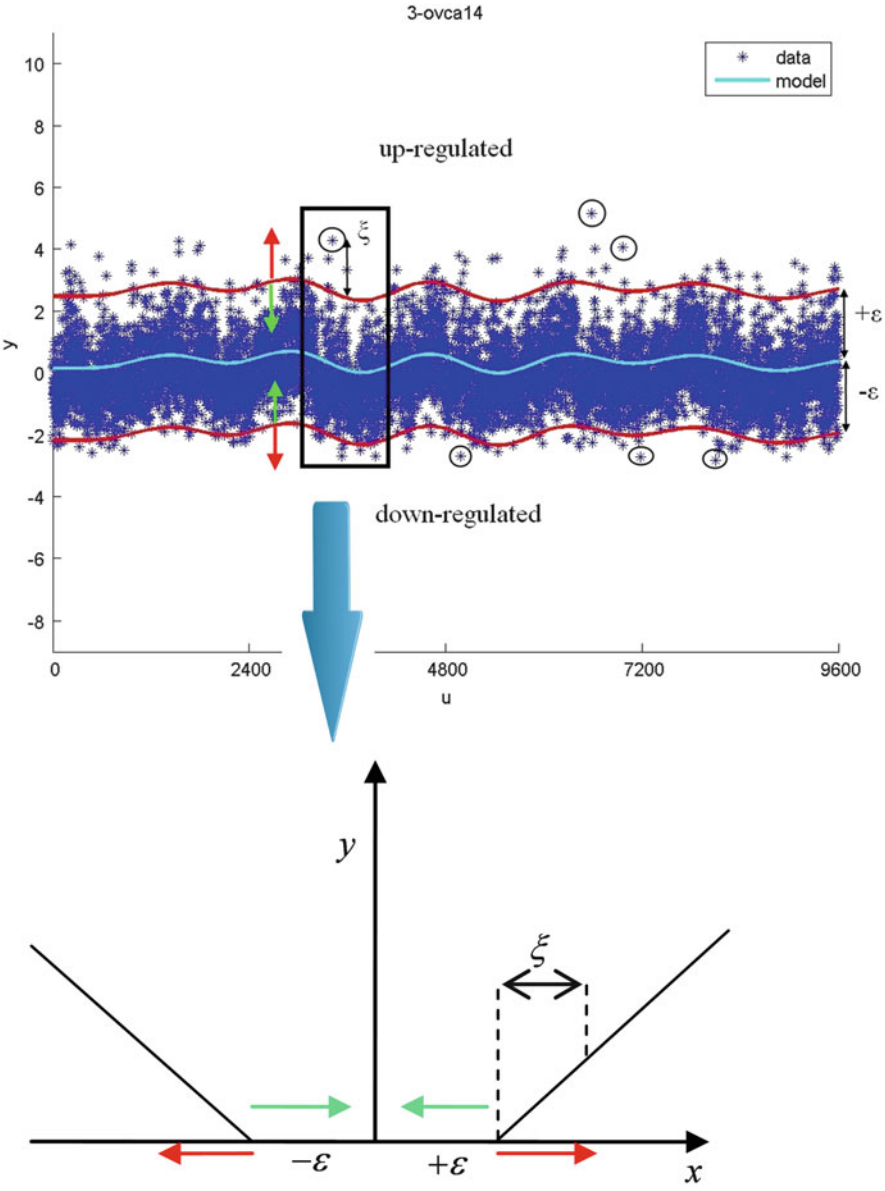

Fig. 19.4 The soft margin loss setting for SVM

19.4 Experiment Results

In this study, the number of experiment sample is 41 and the number of genes is 9,600. Tissues were applied in the current study that included 6 normal, 13 benign OVT, 7 OVCAI, and 15 OVCAIII. Figure 19.5a and b shows the microarray dataset used the proposed approach with different feature genes under fourfold and LOO cross-validations, respectively.

In general, more genes in microarray don't guarantee to get greater prediction classification accuracy as Fig. 19.5. Besides, the prediction classification accuracy hadn't linear relationship with genes number absolute. In Fig. 19.5b, the best prediction classification accuracy with LOO had used two or three genes. Table 19.2 shows the results of experiments that got greater predictive classification accuracy with less than the original gene number, whether fourfold and LOO cross-validation in this chapter.

From the above results, in this chapter we successfully apply SVM for microarray data analysis.

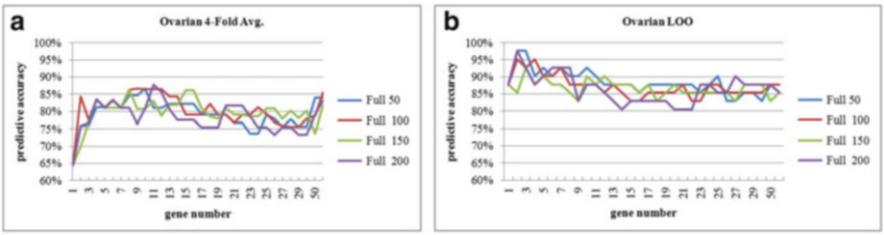

Fig. 19.5 Shown the proposed approach for the ovarian microarray data with different feature genes under (**a**) fourfold cross-validation and (**b**) LOO cross-validation

Table 19.2 The best prediction of classification accuracy of ovarian cancers under different feature genes with fourfold and LOO cross-validation

	Ovarian microarray data			
Gene set	Gene number	Fourfold average (%)	Gene number	LOO (%)
Original	9,600	83.16	9,600	82.93
Full 50	10	86.70	2	97.56
Full 100	9	86.70	2	95.12
Full 150	8	86.29	3	92.68
Full 200	11	87.71	2	97.56

19.5 Conclusions

It is difficult to find out the feature genes for cancer research. Additionally, the cost of disease detection has a relation with the number of genes in microarray. Hence, in this chapter, the proposed approach can reduce the number of genes after epsilon-SVR analysis. Also the simulation results revealed that higher prediction accuracy with less than the original gene number. That means the proposed approach can be effectively applied to selecting feature genes and prediction from microarray data with lower cost.

Acknowledgment The authors wish to thank that this work was supported by National Science Council Under Grant NSC 95-2221-E-150-085, NSC 101-2221-E-150-048-MY2.

References

1. Peterson, C., Ringnér, M.: Analyzing tumor gene expression profiles. Artif. Intell. Med. **28**, 59–74 (2003)
2. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics **16**, 906–914 (2000)
3. Zhang, C., Lu, X., Zhang, X.: Significance of gene ranking for classification of microarray samples. IEEE/ACM Trans. Comput. Biol. Bioinf. **3**, 31–320 (2006)
4. Tang, Y., Zhang, Y.Q., Huang, Z.: Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM Trans. Comput. Biol. Bioinf. **4**, 365–381 (2007)
5. Kung, S.Y., Mak, M.W.: Feature selection for self-supervised classification with applications to microarray and sequence data. IEEE J. Sel. Top. Signal Process. **2**, 297–309 (2008)
6. Leung, Y., Hung, Y.: A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. IEEE/ACM Trans. Comput. Biol. Bioinf. **7**, 108–117 (2010)
7. Lee, C.P., Leu, Y.: A novel hybrid feature selection method for microarray data analysis. Appl. Soft Comput. **11**, 20–213 (2011)
8. Jeng, J.T., Lee, T.T., Lee Y.C.: Classification of ovarian cancer based on intelligent systems with microarray data. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 1053–1058 (2005)
9. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)