The Dependence of SVM-RFE Gene Selection on Microarray Data Noise: A Simulation Study

Elizabeth Tapia[1−2], Pilar Bulacio[1], and Laura Angelone[1−2]

[1] CIFASIS-Conicet Institute, Bv. 27 de Febrero 210 Bis, Rosario, Argentina
[2] Facultad de Cs. Exactas e Ingeniería, Riobamba 245 Bis, Rosario, Argentina

Abstract. A simulation approach to study the dependence of SVM-RFE gene selection with respect to the signal-to-noise ratio of microarray datasets is presented. It is revealed that SVM-RFE gene selection depends on both the signal-to-noise ratio and the policy of gene elimination. Specifically, for SVM-RFE implementations removing a constant fraction of genes per step, smaller signal-to-noise ratios lead to the selection of smaller sets of genes with lower rates of false discoveries. Conversely, for the native SVM-RFE implementation removing one gene per step, larger sets of genes with higher rates of false discoveries, are selected. We conclude that one should be very careful with SVM-RFE gene selection conclusions on microarray data.

1 Introduction

The design of effective classifiers for microarray data entails the selection of smaller sets of relevant genes [1]. However, efficient gene selection from microarray data is strongly limited by two practical constraints: the unusual dimensionality of the data, i.e., the number of genes greatly exceeding the number of samples, and the measurement of hybridization intensities, which are subject to many sources of systematic and random errors [2–4], i.e., intensity changes may not necessarily involve gene expression change. Although many methods of data preprocessing aim to minimize systematic errors in gene-expression, e.g., normalization methods [5], the reduction of random errors needs the availability of an adequate number of replicates, which cannot be satisfied in most biological-research projects. Consequently, residual random fluctuations still remain in microarray data after the removal of systematic errors. This observation raises the question about the dependence between the gene selection results and the ratio of the signal (desired biological variables) and such residual noise, hereafter called signal-to-noise ratio.

To gain insight into gene selection on noisy microarray datasets, a simulation study with focus on SVM-RFE (Support Vector Machine based Recursive Feature Elimination) [6] was performed. Briefly, SVM-RFE gene selection relies on the weight of support vectors. However, as noted by [7], weights of support vectors are not robust against outliers and noises in training data. Hence, SVM-RFE gene selection with rather low ratio of signal-to-noise may be strongly degraded.

XXXIV Conferencia Latinoamericana de Informática (CLEI 2008), pp. 769-774, 2008.

Recently, [8] showed that SVM-RFE allowing the elimination of constant fraction of the remaining genes per selection step could achieve better classification performance than the elimination of a single gene per step (native SVM-RFE). This result contradicts the widespread belief that costly native SVM-RFE gene selection will be always rewarded by optimal classification performance. We hypothesize that a robust policy of gene removal for SVM-RFE has to take into account the signal-to-noise relationship.

In this paper, SVM-RFE gene selection on microarray datasets with different signal-to-noise ratios under multiple policies of gene removal is evaluated. To characterize the signal-to-noise ratio of microarray datasets, simulated microarray datasets constructed from real ones are used. Experiments evaluate the behavior of the classification error, the number of selected genes, and the rate of false discoveries.

2 Methods

2.1 Simulated microarray datasets

Synthetic microarray datasets of well-known signal-to-noise ratios were generated. Aiming to avoid the setting of large number of parameters [9], we generate microarray data numerically starting from the number of expressed genes and the interval of treatment effect. Hence, similarly to [10], our approach starts with a set of irrelevant genes to any sample classification task but showing random fluctuations because of the presence of technical and biological noise. Aiming to mimic real microarray data, the expressions of genes were randomly sampled (with replacement) from eight self-self hybridization experiments used in the labeling study [11]. A small fraction of these irrelevant genes was then selected for the artificial induction of differential expression based on a model of main effects. As a result, the expression of a relevant gene for sample classification was constructed by the addition of a treatment effect proportional to its natural variability, which was estimated by its standard deviation in self-self hybridization experiments. The treatment effect was assumed uniformly distributed in the interval $[Min, Max]$.

Concerning the conjecture that the higher the level of gene expression w.r.t. the noise, the easier the gene selection task, two extreme levels of signal-to-noise ratio were analyzed: low-level modeled by $Min = 0.5$ and $Max = 1$ and high-level modeled by $Min = 2$ and $Max = 4$. These two levels of differential gene expression respectively characterize *low* and *high* signal-to-noise ratio microarray datasets. Our experiments were based on binary microarray datasets of $n = 20$ samples (10 samples per class), and $p = 200$ genes among which just 10% were differentially expressed ($p^* = 20$).

2.2 Variants of SVM-RFE gene selection

SVM-RFE gene selection requires the specification of a filter out factor f modeling the number of genes to be removed at each step. In this paper we follow

the notation introduced in [8]: $f = -1$ identifies the elimination of a single gene per step (native SVM-RFE), and $0 < f < 1$ entails the elimination of a constant fraction f of the remaining genes per step. Since constructing a single SVM using n training examples takes $O\left(n^2 \cdot p\right)$ in time, implementing SVM-RFE with $f = -1$ will take $O\left(n^2 \cdot p^2 \cdot log_2 p\right)$, which can be prohibitive when either n or p are rather large. For the microarray data case, the condition $p \gg n$ holds. So, SVM-RFE implementations allowing the elimination of a constant fraction of the remaining genes per step ($0 < f < 1$) are practically used.

It is widely believed [6] that optimal SVM-RFE classification accuracy requires $f = -1$. However, recent results [8] suggest that this assumption is not completely true. Aiming at disentangling the impact of the filtering out factor f in SVM-RFE gene selection, a carefully controlled simulation study was designed.

2.3 Simulation study

SVM-RFE gene selection based on linear SVM classifiers with default constant complexity $C_{SVM} = 1$ was evaluated on synthetic microarray datasets. The generalization performance of resulting SVM classifiers was estimated by 10 Fold Cross Validation (10-FoldCV) error. At each fold, optimum gene selection at default $C_{SVM} = 1.0$ was followed by C_{SVM} optimization over the grid $\left\{2^{-5}, \ldots, 1, \ldots, 2^5\right\}$.

To avoid gene selection biases, both the gene selection and the C_{SVM} optimization were performed using external 10-FoldCV error loops [12]. The complete optimization process was performed by means of the R packages MCREstimate [13] and SVM-RFE [12]. The complete experimental protocol was repeated on four different simulated datasets. Hence, reported statistics rely on forty train-test partitions (4 x 10-FoldCV).

In addition to the evaluation of the 10-Fold CV error, SVM-RFE gene selection was evaluated by means of the mean number of selected genes, and the rate of false discoveries occurring in lists of frequently selected genes, defined as those genes selected at least 90% of times on the 10 FoldCV error evaluations).

Under this baseline, the robustness of SVM-RFE gene selection against noise under different settings of the parameter $f \in \{-1, 0.2, 0.35, 0.5\}$ was evaluated. We tested the hypothesis that misleading results about SVM-RFE gene selection can be explained by the SVM sensitivity to noise and outliers. If this were true, good SVM-RFE gene selection on high signal-to-noise microarray datasets would require $f = -1$, but larger $f's$ would be better to prevent the risk of overfitting on rather low signal-to-noise ratio microarray datasets. This hypothesis was confirmed by simulation results.

3 Results and discussion

Table 1 shows the SVM-RFE gene selection performance on synthetic microarray data. Let FDR be the rate of false discoveries in lists of selected genes of size NGenes.

Table 1. Signal-to-noise vs. f relationship. "Error" is the mean 10-Fold CV error. "Genes" is the mean number of selected genes in 10-Fold CV error evaluations. "NGenes" is number of genes selected at least once. "NGenes90" is number of genes selected at least nine times, i.e., frequently selected genes. "FDR" and "FDR90" are respectively the fractions of false discoveries in raw sets of selected and frequently selected genes. "TE" and "TE90" are respectively the number of truly expressed genes reported in raw sets of selected and frequently selected genes. The .(sd) is the standard deviation.

	High Signal-to-Noise Ratio			
f	**-1**	**0.2**	**0.35**	**0.5**
Error (sd)	0.17 (0.06)	0.012 (0.02)	0.012 (0.02)	0.025 (0.03)
Genes (sd)	1.37 (0.25)	117.8 (20.14)	78.72 (29.67)	43.87 (21.87)
NGenes (sd)	5 (1.4)	188.5 (6.45)	155.25 (24.14)	94.5 (43.55)
FDR (sd)	0 (0)	0.90 (0.007)	0.88 (0.017)	0.77 (0.09)
TE (sd)	5 (1.4)	19 (1.4)	18.75 (1.25)	18.5 (1.73)
NGenes90 (sd)	0 (0)	61.5 (29.89)	39.25 (18.19)	13.25 (7.09)
FDR90 (sd)	NA	0.88 (0.02)	0.89 (0.05)	0.80 (0.12)
TE90 (sd)	NA	6.75 (2.06)	4.25 (3.2)	2.25 (1.5)
	Low Signal-to-Noise Ratio			
f	**-1**	**0.2**	**0.35**	**0.5**
Error (sd)	0.24 (0,05)	0.07 (0.05)	0.14 (0.07)	0.11 (0.07)
Genes (sd)	5.85 (2.17)	55.85 (18.15)	31.1 (14.15)	17.42 (6.72)
NGenes (sd)	16.5 (6.35)	127.5 (38.90)	72.25(29.90)	39.75 (21.50)
FDR (sd)	0.24 (0.08)	0.84 (0.06)	0.73 (0.08)	0.46 (0.32)
TE (sd)	12.25 (4.35)	18.75 (1.26)	17.5 (1.9)	16.75 (2.2)
NGenes90 (sd)	0 (0)	24 (10.23)	14 (5.29)	8 (4.08)
FDR90 (sd)	NA	0.85 (0.08)	0.72 (0.16)	0.46 (0.37)
TE90 (sd)	NA	3 (1.63)	3.5 (1.91)	3.75 (2.75)

Regarding microarray datasets with rather high signal-to-noise ratios, middle values of f seem to be more appropriated: $f = \{0.2, 0.35\}$ lead to better relationship between the number of truly expressed genes (TE) and the classification error (Error) than $f = -1$. Furthermore, for microarray datasets with rather low signal-to-noise ratios, even $f = 0.2$ seems to be more appropriate than $f = -1$. In other words, optimal gene SVM-RFE gene selection may require some knowledge about the inherent quality of the microarray dataset at hand. This knowledge may be used to select the more appropriate f parameter, which may lead to important savings in the computational processing of SVM-RFE gene selection.

Regarding the stability of SVM-RFE gene selection, lists of frequently selected genes were further screened to characterize their power in the identification of sets of genes of biological relevance. Without loss of generality, we focused on genes with a frequency of selection at least 90%. Let NGenes90 be the size of lists of frequently selected genes. The native SVM-RFE implementation

($f = -1$) attained unstable sets of genes on both low and high signal-to-noise ratios (NGenes90=0). But when using middle values of f, near half of the selected genes were indeed frequently selected, i.e, $NGenes90 \approx 0.5 \cdot Genes$.

Regarding the consistency of SVM-RFE gene selection, for $f \neq -1$, decreasingly smaller sets of *truly* expressed genes (TE) are also selected. This reduction is consistent with the selection of smaller sets of genes (Genes). Note, however, that for $f \neq -1$, an appreciate number of truly expressed genes remain frequently selected (TE90). Hence, middle values of f may also be more appropriate to guarantee the replicability of research results based on SVM-RFE gene selection.

Overall, our results suggest that more reliable SVM-RFE models of gene selection may be attained by examining a range of f parameters.

4 Conclusions

How SVM-RFE gene selection is affected by the noise of microarray data? An answer is given within our simulation study on synthetic microarrays with different signal-to-noise ratios and diverse policies of gene removal. As a result, meticulous genes elimination (parameter $f = -1$) seems to be appropriate for microarray datasets with high signal-to-noise ratios. As a result, larger $f's$ may be safely used for challenging microarray data classification problems. Besides the classification error, microarray data classifiers are judged also by their ability to predict with small and stable sets of genes. For the native SVM-RFE gene selection, the stability feature is not achieved. On the other hand, rough SVM-RFE gene selection seems to be stable even for rather low signal-to-noise ratio microarray datasets. Finally, relevant genes to a biological problem can be extracted from lists of frequently selected genes as well as the model stability. At the end, we can conclude that fixed policies of gene elimination cannot achieve good classification performance over a wide spectrum of signal-to-noise ratios.

Acknowledgment

The authors would like to thank Javier De Las Rivas, CSIC-Salamanca, Spain, for providing access to computational resources and Julio Di Rienzo, Universidad Nac. de Córdoba, Facultad de Agronomía, Argentina, for providing the R script for synthetic microarray data generation. ET's, PB's, and LA's work was supported by project PICT No. 02226- 2006, SECYT, Argentina.

References

1. Stolovitzky, G.: Gene selection in microarray data: the elephant, the blind men and our algorithms. Curr Opin Struct Biol **13** (2003) 370–376
2. Raser, J., O'Shea, E.K.: Noise in gene expression: Origins, consequences, and control. Science **309** (2005) 2010–2013

3. Tu, Y., Stolovitzky, G., Klein, U.: Quantitative noise analysis for gene expression microarray experiments. Proceedings of the National Academy of Sciences **99** (2002) 14031–14036
4. Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S.B., Bassett, D.E.: Rosetta error model for gene expression analysis. Bioinformatics **22** (2006) 1111–1121
5. Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., Speed, T.P.: Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res **30** (2002)
6. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46** (2002) 389–422
7. Boser, B.E., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Computational Learing Theory. (1992) 144–152
8. 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. Bioinformatics **4** (2007) 365–381
9. Balagurunathan, Y., Dougherty, E., Chen, Y., Bittner, M., Trent, J.: Simulation of cdna microarrays via a parameterized random signal model. Biomedical Optics **7** (2002) 507–523
10. Yeung, K., Bumgarner, R.: Multiclass classification of microarray data with repeated measurements: application to cancer. Genome Biol **4** (2003) R83
11. Manduchi, E., Scearce, L.M., Brestelli, J.E., Grant, G.R., Kaestner, K.H., Stoeckert, C.J.: Comparison of different labeling methods for two-channel high-density microarray experiments. Physiol Genomics **10** (2002) 169–179
12. Ambroise, C., Mclachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA **99** (2002) 6562–6566
13. Ruschhaupt, M., Mansmann, U., Warnat, P., Huber, W., Benner, A.: Misclassification error estimation with cross-validation, MCRestimate. (2004)