On the Effectiveness of Gene Selection for Microarray Classification Methods

Zhongwei Zhang[1], Jiuyong Li[2], Hong Hu[3], and Hong Zhou[4]

[1] Department of Mathematics and Computing, University of Southern Queensland
QLD 4350, Australia
zhongwei@usq.edu.au
[2] School of Computer and Information Science
University of South Australia,
Mawson Lakes, Adelaide, SA 5001, Australia
Jiuyong.Li@unisa.edu.au
[3] Planning and Quality Office, University of Southern Queensland
QLD 4350, Australia
huhong@usq.edu.au
[4] Faculty of Engineering, University of Southern Queensland
QLD 4350, Australia
hzhou@usq.edu.au

Abstract. Microarray data usually contains a high level of noisy gene data, the noisy gene data include incorrect, noise and irrelevant genes. Before Microarray data classification takes place, it is desirable to eliminate as much noisy data as possible. An approach to improving the accuracy and efficiency of Microarray data classification is to make a small selection from the large volume of high dimensional gene expression dataset. An effective gene selection helps to clean up the existing Microarray data and therefore the quality of Microarray data has been improved. In this paper, we study the effectiveness of the gene selection technology for Microarray classification methods. We have conducted some experiments on the effectiveness of gene selection for Microarray classification methods such as two benchmark algorithms: SVMs and C4.5. We observed that although in general the performance of SVMs and C4.5 are improved by using the preprocessed datasets rather than the original data sets in terms of accuracy and efficiency, while an inappropriate choice of gene data can only be detrimental to the power of prediction. Our results also implied that with preprocessing, the number of genes selected affects the classification accuracy.

1 Introduction

Gene selection technology has been widely used by many researchers in the past decades to select the most effective genes from high dimensional Microarray data. The Microarray gene data acquired from Microarray technology is quite different than that from the normal relational databases. Normal relational databases contain a small number of attributes and a large number of samples. In contrast, gene expression Microarray data usually contains a very large number of

attributes but a small number of usable samples. With a large number of genes, it is absolutely desirable to have a large number of samples accordingly in order to build reliable Microarray classification models. However, the reality is that for most Microarray experiments, a limited number of samples are available due to the huge cost of producing such Microarray data and other factors, such as privacy and availability. As an example, for cancer Microarray data, the number of samples is usually less than 200.

In short, high dimensionality renders many classification methods not applicable for analyzing raw gene Microarray data. Furthermore, high dimension Microarray data with noisy attributes leads to unreliable and low accuracy analysis results. Consequently, reducing irrelevant and removing noise gene expression values from the original Microarray data are crucial for applying classification algorithms to analyze gene expression Microarray data.

Many researches have shown that gene selection can improve the performance of Microarray classification [7, 20, 26–28]. But these research haven't answered the question: *Can good gene selection methods enhance the prediction performance of all types of Microarray classification methods, wrapper classification methods in particular?*

This paper is organized as follows. In the preceding section, we identify problems in gene expression Microarray data classification and highlight the importance of gene selection for gene expression Microarray data. In Section 2, we review a number of gene selection methods. In Section 3, we present the design of methods for comparing the accuracy of SVMs and C4.5 using different gene selection methods. In Section 4, we test four different gene selection methods with six data sets. In Section 5, we present a discussion of the results. In Section 6, we conclude the paper.

2 Gene selection methods

To deal with the problems caused by high dimensionality and noisy Microarray gene data, a preprocessing phase should be introduced to reduce the noise and irrelevant genes before the Microarray data classifications are applied. As a preprocessing method of Microarray data classification, gene selection is a very effective way for eliminating the noisy genes. In essence, gene selection aims to select a relatively small set of genes from a high dimensional gene expression data set. Gene selection helps to clean up the existing Microarray data and therefore improve the quality of Microarray data. In other words, removing irrelevant and noisy genes is helpful for improving the accuracy of Microarray data classification. The resultant classification models of Microarray gene data would therefore better characterize the true relationships among genes and hence be easier to be interpreted by biologists. Arguably, a good gene selection method not only increases the accuracy of classification through the improvement of the Microarray data quality, but also speeds up the classification process through the cutdown of high dimensionality.

Based on the dependency on classification algorithms, gene selection methods can be roughly divided into wrapper and filter methods [15]. A filter method performs gene selection independently from a classification method. It preprocesses a Microarray data set before the data set is used for classification analysis. Some filter gene selection methods are: ranking gene selection methods [22], and information gain gene selection method [21], Markov blanket-embedded genetic algorithm for gene selection [32], and so on. One-gene-at-a-time filter methods, such as ranking [22], signal-to-noise [27], information gain [24], are fast and scalable but do not take the relationships between genes into account. Some genes among the selected genes may have similar expression levels among classes, and they are redundant since no additional information is gained for classification algorithms by keeping them all in the dataset. To this end, Koller and Sahami [17] developed an optimal gene selection method called *Markov blanket filtering* which models feature dependencies and can eliminate redundant genes. Further to this method, Yu and Liu [31] proposed the Redundancy Based Filter(RBF) method, which is able to deal with redundant problems. Favorable results have been achieved.

In contrast, a wrapper method embeds a gene selection method within a classification algorithm. An example of a wrapper method is SVMs [11], which uses a recursive feature elimination(RFE) approach to eliminate the features iteratively in a greedy fashion until the largest margin of separation is reached. Wrapper methods are not as efficient as filter methods due to the fact that they usually run on the original high dimensional Microarray dataset. However, Kohavi and John [15] discovered that wrapper methods could significantly improve the accuracy of classification algorithms over filter methods. This discovery indicates that the performance of a classification algorithm is largely dependent on the chosen gene selection method. Nevertheless, no single gene selection method can universally improve the performance of classification algorithms in terms of efficiency and accuracy.

In Section 3, we design some experiments to investigate the dependency between gene selection methods and Microarray data classification methods.

3 Experimental design and methodology

Our approach is to use different existing gene selection methods to preprocess Microarray data for classification. We have carried out our experiments by comparing with benchmark algorithms SVMs and C4.5. Note that this choice is based on the following considerations.

Consideration of benchmark systems: For years, SVMs and C4.5 have been regarded as benchmark classification algorithms. SVMs was proposed by Cottes and Vapnik [5] in 1995. It has been one of the most influential classification algorithms. SVMs has been applied to many domains, for example, text categorization [14], image classification [23], cancer classification [9, 2]. SVMs can easily deal with high dimensional data sets with a wrapper gene selection method.

SVMs also can achieve a higher performance compared to most existing classification algorithms.

Considering of wrapper methods: SVMs and C4.5 are not only benchmark classification systems, but each of them contains a wrapper gene selection method. SVMs uses a recursive feature elimination(RFE) approach to eliminate the features iteratively in a greedy fashion until the largest margin of separation is reached. Decision tree method can also be treated as a gene selection method. It selects the gene with the highest information gain at each step and all selected genes appear in the decision tree.

A ranking method identifies one gene at a time with differentially expressed levels among predefined classes and puts all genes in decreasing order. After a specified significance expressed level or number of genes is selected, the genes lower than the significance level or given number of genes are filtered out. The advantages of these methods is that they are intuitive, simple and easy to implement. In this study, we choose and implement four popular ranking methods collected by Cho and Won [3], namely Signal-to-Noise ratio (SNR), correlation coefficient (CC), Euclidean (EU) and Cosine (CO) ranking methods.

To evaluate the performance of different gene selection methods, six datasets from Kent Ridge Biological Data Set Repository [19] were selected. These data sets were collected from some influential journal papers, namely the breast cancer, lung cancer, Leukemia, lymphoma, colon and prostate data sets 3. Each Microarray dataset is described by the following parameters. (1) Genes: the number of genes or attributes (2) Class: the number of classes, (3) Record: the number of samples in the dataset

Table 1. Gene expression Microarray data sets

	Dataset name	Genes	Class	Sample
1	Breast Cancer	24481	2	97
2	Lung Cancer	12533	2	181
3	Lymphoma	4026	2	47
4	Leukemia	7129	2	72
5	Colon	2000	2	62
6	Prostate	12600	2	21

During the gene expression Microarray data preprocessing stage, we define the number of selected genes as 20, 50 and 100 and 200 for all filter gene selection methods. In our experiments, a tenfold cross-validation method is also carried out for each classification method to test its accuracy.

4 Experimental results and discussions

Figure 1 - 12 show the detailed results for SVMs and C4.5 tested on six different datasets preprocessed by four different filter gene selection methods.

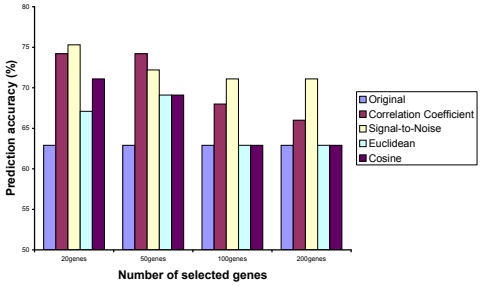

Fig. 1. C4.5 tested on Breast cancer dataset

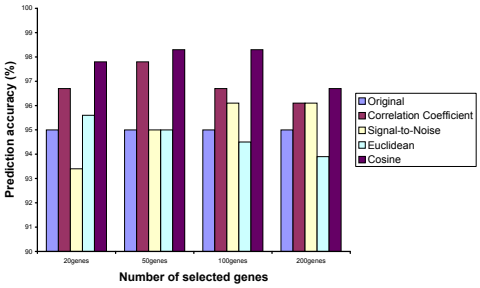

Fig. 2. C4.5 tested on lung cancer dataset

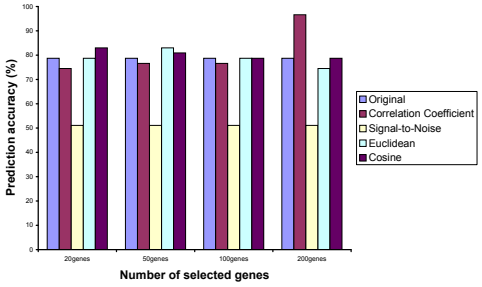

Fig. 3. C4.5 tested on Lymphoma dataset

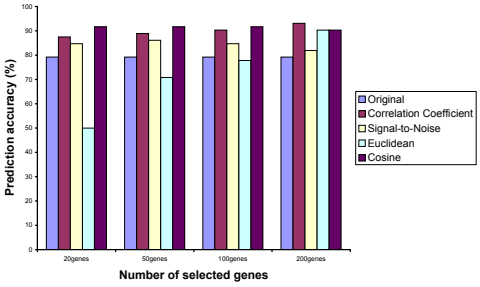

Fig. 4. C4.5 tested on Leukemia dataset

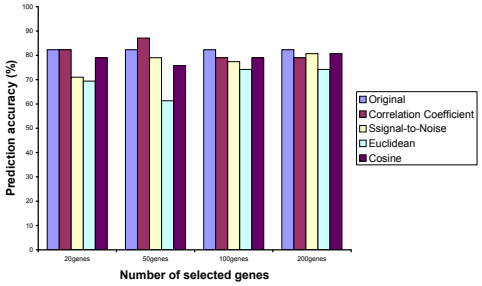

Fig. 5. C4.5 tested on Colon dataset

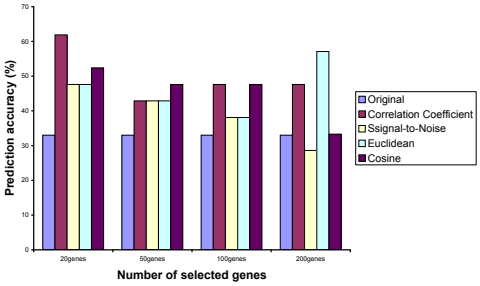

Fig. 6. C4.5 tested on Prostate dataset

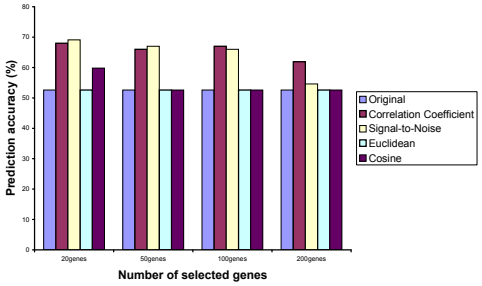

Fig. 7. SVM tested on Breast cancer dataset

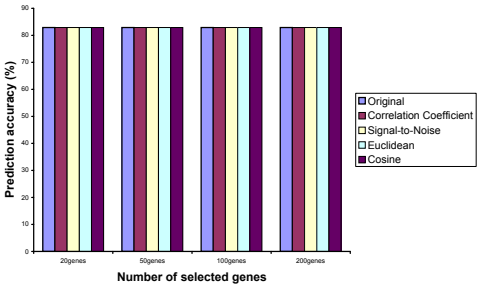

Fig. 8. SVM tested on lung cancer dataset

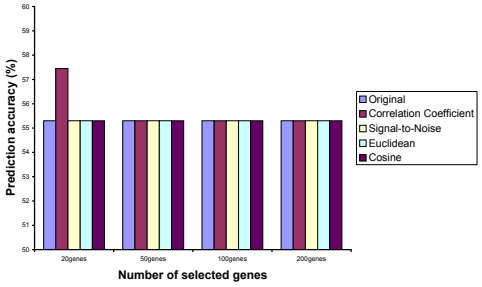

Fig. 9. SVM tested on Lymphoma data set

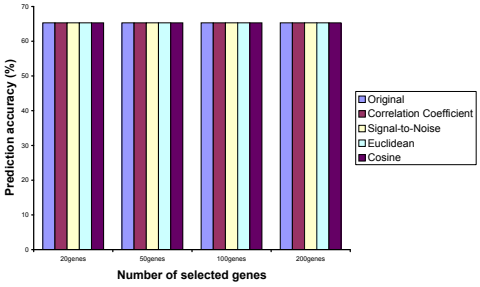

Fig. 10. SVM tested on Leukemia data set

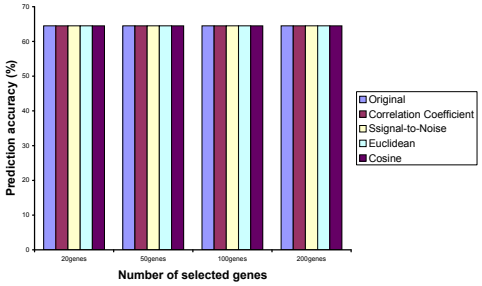

Fig. 11. SVM tested on Colon dataset

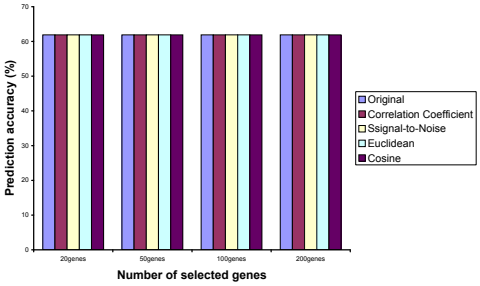

Fig. 12. SVM tested on Prostate dataset

From these experimental results, we make the following observations.

When Microarray data sets are preprocessed, SVMs improves its prediction accuracy on Breast Cancer and Lymphoma data sets only Signal-to-Noise and Correlation coefficient methods performed best and improved the accuracy up to 16.5% and 15.5% respectively on Cancer data. The Cosine method also improved the accuracy by up to 7.2% on Cancer data. On the Lymphoma data set, the Correlation coefficient method is the only method which improved the accuracy performance over original data set by up to 2.2% while other methods were not able to improve the accuracy. None of the gene selection methods improved nor decreased the prediction accuracy based on Lung Cancer, Leukemia, Colon and Prostate data sets with 200, 100, 50 and 20 genes. Instead, the accuracy performance is kept unchanged.

The performance of C4.5 improves its prediction accuracy by up to 28.6% . Among the four gene selection methods, Correlation coefficient is the most effective preprocessing method with an improvement of accuracy up to 7.6% on average, followed by Cosine 7.3%, and Signal-to-Noise 6.0%. The Signal-to-Noise gene selection method performed consistently better on Breast Cancer and Leukemia data sets with improved accuracy by up to 12.4%, but failed on the other cancer data sets. Euclidean in contrast performed worst among the compared methods, decreasing the accuracy performance on all provided cancer data sets except Breast and Prostate by up to 18.7%.

The experimental results show that with preprocessing, the number of genes selected has an affect on some classification methods in terms of performance accuracy. In the figures for C4.5, the highest accuracy for all cancer data sets except the prostate cancer data set are based on 50 genes; while the highest accuracy for prostate cancer data set is based on 20 genes. The overall performance is better when data sets contain 50. However, the number of genes selected has little impact on the performance of SVMs.

5 Discussion of experimental results

In this section, we discuss the implication of gene selection methods upon the classification methods.

The results indicate that gene selection improves the performance of classification methods in general. Using a suitable gene selection method with C4.5 increases the accuracy performance of C4.5 dramatically. For SVMs, its performance remained unchanged unless a very small size of genes was selected. Moreover, gene selection does not decrease the accuracy performance of SVMs. This result ensures that we can reduce the number of genes to a smaller size without hurting the accuracy performance of classification. This is very helpful for noisy Microarray data classification as most irrelevant genes in Microarray data classification would be reduced. It increases the performance of classification significantly in terms of speeding up the efficiency of Microarray data classification.

These results indicate that not all gene selection methods help the performance of Microarray classification methods in terms of improving the prediction accuracy of classification. Their performance depends on which Microarray classification method they are combined with. For C4.5, with the help of some gene selection, such as the Correlation coefficient method, the accuracy performance improved significantly. The Signal-to-Noise method generated mixed results combined with C4.5; while the Euclidean method is not a suitable gene selection method for C4.5 as it failed to improved the accuracy performance of C4.5 on most data sets. So to apply gene selection to C4.5, we have to seriously consider which gene selection algorithm to use to achieve maximum improvement. With SVMs, only the Correlation coefficient method managed to improve the accuracy performance on up to two data sets.

Gene selection may have little impact on some classification methods. The figures show that SVMs is insensitive to the gene selection methods used and hence data preprocessing does not increase its performance in most cases. This indicates that the SVMs classification method can initially handle noise data very well. Moreover, it would require little effort to select a gene selection method for SVMs.

The observations indicate that a data set with less genes or attributes does not necessarily guarantee the highest prediction accuracy. The number of genes selected by a preprocessing method should not be too small. At this stage, the objective of gene selection is just to eliminate irrelevant and noise genes. However, less informative genes can sometimes enhance the power of classification if they are co-related with the most informative genes. If the number of genes has been eliminated too harshly, it can also decrease the performance of the classification. So during the preprocessing, we need to make sure that a reasonable number of genes are left for classification.

Those results remind us that when selecting the gene selection method for data preprocessing, we must consider which classification method the gene selection is for. For example, if we select SVMs as a classification algorithm, then the Correlation coefficient or Signal-to-Noise gene selection methods are better for data preprocessing. An inappropriate choice can only harm the power of classification prediction.

6 Conclusions

In this paper, we have looked into the gene selection technique to improve the quality of Microarray data sets on Microarray data classification methods: SVMs and C4.5, which themselves contain a wrapped method. We observed that although in general the performance of SVMs and C4.5 are improved by using the preprocessed datasets rather than original data sets in terms of accuracy and efficiency, not all gene selection methods help improve the performance of classification. The rule-of-thumb is that some gene selection methods are suitable

for some specific classification algorithms. For example, if we select SVMs as the classification algorithm, then a Correlation coefficient or Signal-to-Noise gene selection method is better for data preprocessing. On the contrary, an inappropriate choice can only harm the power of prediction. Our results also implied that with preprocessing, the number of genes selected affects the classification accuracy.

References

1. R. Blanco, P. Larrañaga, I. Inza, and B. Sierra. Gene selection for cancer classification using wrapper approaches. *IJPRAI*, 18(8):1373–1390, 2004.
2. M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Jr, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using suport vector machines. In *Proc. Natl. Acad. Sci.*, volume 97, pages 262–267, 2000.
3. S.-B. Cho and H.-H. Won. Machine learning in dna microarray analysis for cancer classification. In *CRPITS '19: Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003*, pages 189–198, Darlinghurst, Australia, Australia, 2003. Australian Computer Society, Inc.
4. S.-B. Cho and H.-H. Won. Cancer classification using ensemble of neural networks with multiple significant gene subsets. *Appl. Intell*, 26(3):243–250, 2007.
5. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
6. M. Dash and H. Liu. Feature selection for classification. *Intell. Data Anal*, 1(1-4):131–156, 1997.
7. C. H. Q. Ding. Unsupervised feature selection via two-way ordering in gene expression analysis. *Bioinformatics*, 19(10):1259–1266, 2003.
8. C. H. Q. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinformatics and Computational Biology*, 3(2):185–206, 2005.
9. T. S. Furey, N. Christianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Hauessler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.
10. C. Furlanello, M. Serafini, S. Merler, and G. Jurman. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics*, 4:54, 2003.
11. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
12. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
13. S.-Y. Ho, C.-C. Lee, H.-M. Chen, and H.-L. Huang. Efficient gene selection for classification of microarray data. In *IEEE Congress on Evolutionary Computation*, pages 1753–1760. IEEE, 2005.
14. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142. Springer Verlag, Heidelberg, DE, 1998.
15. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

16. D. Koller and M. Sahami. Toward optimal feature selection. In *ICML*, pages 284–292, 1996.
17. D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.
18. Z.-J. Lee. An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer. *Artificial Intelligence in Medicine*, 42(1):81–93, 2008.
19. J. Li and H. Liu. Kent ridge bio-medical data set repository. http://sdmc.lit.org.sg/gedatasets/datasets.html, 2002.
20. S. Li, X. Wu, and X. Hu. Gene selection using genetic algorithm and support vectors machines. *Soft Comput*, 12(7):693–698, 2008.
21. X. Liu, A. Krishnan, and A. Mondry. An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics*, 6:76, 2005.
22. S. Mukkamala, Q. Liu, R. Veeraghattam, and A. H. Sung. Feature selection and ranking of key genes for tumor classification: Using microarray gene expression data. In L. Rutkowski, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, editors, *ICAISC*, volume 4029 of *Lecture Notes in Computer Science*, pages 951–961. Springer, 2006.
23. E. Osuna, R. Freund, and F. Girosi. Training support vector machines:an application to face detection. In *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997.
24. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.
25. Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.
26. M. Song and S. Rajasekaran. A greedy correlation-incorporated SVM-based algorithm for gene selection. In *AINA Workshops (1)*, pages 657–661. IEEE Computer Society, 2007.
27. T.R.Golub, D.K.Slonim, P. Tamayo, and et.al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
28. L. V. Veer, H. Dai, M. V. de Vijver, and et.al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.
29. X. wen Chen. Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines. In *CSB*, pages 504–505. IEEE Computer Society, 2003.
30. K. Yendrapalli, R. B. Basnet, S. Mukkamala, and A. H. Sung. Gene selection for tumor classification using microarray gene expression data. In S. I. Ao, L. Gelman, D. W. L. Hukins, A. Hunter, and A. M. Korsunsky, editors, *World Congress on Engineering*, Lecture Notes in Engineering and Computer Science, pages 290–295. Newswood Limited, 2007.
31. L. Yu and H. Liu. Redundancy based feature selection for microarray data. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA,*, pages 737–742, 2004.
32. Z. Zhu, Y.-S. Ong, and M. Dash. Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, 40(11):3236–3248, 2007.