Gene Ranking from Microarray Data for Cancer Classification–A Machine Learning Approach

Roberto Ruiz[1], Beatriz Pontes[1], Raúl Giráldez[2], and Jesús S. Aguilar–Ruiz[2]

[1] Department of Computer Science, University of Seville
Avenida Reina Mercedes s/n, 41012 Sevilla, Spain
{rruiz, bepontes}@lsi.us.es

[2] Area of Computer Science, University of Pablo de Olavide
Ctra. de Utrera, km. 1, 41013, Sevilla, Spain
{rgirroj, jsagurui}@upo.es

Abstract. Traditional gene selection methods often select the top–ranked genes according to their individual discriminative power. We propose to apply feature evaluation measure broadly used in the machine learning field and not so popular in the DNA microarray field. Besides, the application of sequential gene subset selection approaches is included. In our study, we propose some well-known criteria (filters and wrappers) to rank attributes, and a greedy search procedure combined with three subset evaluation measures. Two completely different machine learning classifiers are applied to perform the class prediction. The comparison is performed on two well–known DNA microarray data sets. We notice that most of the top-ranked genes appear in the list of relevant–informative genes detected by previous studies over these data sets.

1 Introduction

The gene expression data are typically organized in microarrays. These are matrices where columns represent genes and rows represent experimental conditions (henceforth samples). Each element in the matrix refers to the expression level of a particular gene under a specific condition.

Analysis of microarray data presents unprecedented opportunities and challenges for data mining in areas such as gene clustering [1], sample clustering and class discovery [1,4], sample classification [4] and gene selection [6,9,16,18]. In this work, we address the gene selection issue under a classification framework. The task is to build a classifier that accurately predicts the classes (diseases or phenotypes) of new unlabeled samples. A typical data set may contain thousand of genes but only small number of samples (often less than two hundred). Theoretically, having more features should give us more discriminating power. However, this can cause several problems: increase computational complexity and cost; too many redundant or irrelevant genes; and degradation of the estimation of the classification error. In addition to reducing noise and improving

* This research was supported by the Spanish Research Agency CICYT under grants TIN2004–00159 and TIN2004-06689C0303.

the accuracy of classification, the selected subsets of genes may have important biological interpretation and may be used for drug target discovery or identifying future possible research directions.

In this work, we carry out a study of the performance that several feature selection methods show with two microarrays: Colon Cancer [1] and Leukemia [4]. Although such methods are widely applied in machine learning area, they are not so popular in the DNA microarray field. The application of sequential gene subset selection approaches is included too. In particular, we used six filter and three wrapper methods to rank attributes, and a greedy search procedure combined with three subset evaluation measures. Two well-known machine learning classifiers (naive Bayes and C4.5 [13]), with completely different approaches to learning, are applied to perform the class prediction. This analysis shows that most of the top-ranked genes appear in the list of relevant–informative genes detected by previous studies over these data sets.

The paper is organized as follows. We introduce feature (gene) selection for classification and related work in the next section. Experimental results are shown in Section 3, and the most interesting conclusions are summarized in Section 4.

2 Feature Selection for Classification

The problem of feature selection received a thorough treatment in pattern recognition and machine learning [12]. The gene expression data sets are problematic in that they contain a large number of genes (features) and thus methods that search over subsets of features can be prohibitively expensive. Moreover, these data sets contain only a small number of samples, so the detection of irrelevant genes can suffer from statistical instabilities. Feature selection is reviewed in two ways according to the evaluation measure: depending on their dependency on mining algorithms or based on the way that features are evaluated.

2.1 Filter and Wrapper Model

Feature selection algorithms designed with different evaluation criteria broadly fall into two categories [12]: the filter model and the wrapper model. The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance, but it also tends to be more computationally expensive than filter model. A hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages [16,18].

As described in [12], some popular criteria are distance, information, dependency, consistency and performance of a classifier measures. A large number of

measures have been proposed for scoring genes in the microarray field: Golub et al. [4] proposed PS (Prediction Strength);Ben-Dor et al. [2] TNoM score (Threshold Number of Misclassification); information gain [16]; t-score [17]; and LDA (Linear Discriminant Analysis), LR (Logistic Regression) and SVM (Support Vector Machine).

2.2 Individual and Subset Evaluation

There exist two major approaches in gene/feature selection from the method's output point of view: feature ranking (FR) and feature subset selection (FSS), depending on the way that features are evaluated. The first one, also called feature weighting [5], assesses individual features and assigns them weights according to their degrees of relevance, while the second one evaluates the goodness of each found feature subset.

In the FR algorithms category, one can expect a ranked list of features which are ordered according to evaluation measures. A subset of features is often selected from the top of a ranking list. A feature is good and thus will be selected if its weight of relevance is greater than a user-specified threshold value, or we can simply select the first k features from the ranked list. This approach is efficient due to its linear time complexity in terms of dimensionality.

In the FSS algorithms category, candidate feature subsets are generated based on a certain search strategy. Each candidate subset is evaluated by a certain evaluation measure and compared with the previous best one with respect to this measure. If a new subset turns out to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Different algorithms address theses issues differently. In [12], a great number of selection methods are categorized. We found different search strategies, namely exhaustive, heuristic and random search, combined with several type of measures to form different algorithms. The time complexity is exponential in terms of data dimensionality for exhaustive search and quadratic for heuristic search. The complexity can be linear to the number of iterations in a random search, but experiments show that in order to find the best feature subset, the number of iterations required is mostly at least quadratic to the number of features [3].

Most popular search methods in machine learning can not be applied to microarray data sets due to the large number of genes. Usually, existing algorithms rank genes according to heir individual relevance or discriminative power to the targeted classes and select top-ranked genes.

Some existing subset evaluation measures in machine learning that have been shown effective in removing both irrelevant and redundant features include the consistency measure [3], the estimated accuracy of a learning algorithm, and the correlation measure [7]. Above-mentioned are the two first, and correlation measure evaluates the goodness of feature subsets based on the hypothesis that good feature subsets contain features highly correlated to the class, yet uncorrelated to each other.

3 Experiments and Results

In this section, a comparison among a group of different filter and wrapper metrics is carried out. Besides, we empirically evaluate the efficiency and effectiveness of three FSS approaches on gene expression microarray data. Descriptions of the two data sets are studied follow.

Colon cancer data set. This data set is a collection of expression measurements from colon biopsy samples reported by Alon et al. [1]. The data set consists of 62 samples of colon epithelial cells. These samples were collected from colon-cancer patients. The ŞtumorŤ biopsies were collected from tumors, and the ŞnormalŤ biopsies were collected from healthy parts of the colons of the same patients. The final assignments of the status of biopsy samples were made by pathological examination. Of the $\approx$ 6000 genes represented in these arrays, 2000 genes were selected based on the confidence in the measured expression levels.

Leukemia data set. This data set is a collection of expression measurements reported by Golub et al. [4]. The data set contains 72 samples. These samples are divided to two variants of leukemia: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL). The source of the gene expression measurements was taken from 63 bone marrow samples and 9 peripheral blood samples. The expression levels of 7129 genes are reported.

The experiments are conducted using the WEKA's implementation of all these existing algorithms[15]. In order to apply some of the measures (ig, cn and c4), the expression values of each gene are discretized previously.

3.1 Classification with the Genes of Highest Scoring Value

In our study, we apply nine well-known criteria to rank attributes, each of them has a long tradition in feature selection and statistics literature. Six of them are filters: information gain (IG), non-linear correlation (CR) and consistency (CN) are mentioned in Section 2.1 (information, dependency and consistency measures), and ReliefF (RL) [10], Soap (SP) [14] and Chi2(CH) [11]. In the three wrapper approaches applied, naive Bayes (WNB), instance-based (WIB) and c4.5 (WC4) classifiers are used to provide the ranked list. For each metric, we construct the classification models with three, five, ten and twenty genes of highest scoring value. Thus, for each ranked-list, the same subset of genes is used to build the two classification models with Bayesian classifier (NB) and C4.5 (C4).

The main contribution of this study is the use of some criteria for ranking genes that have rarely been used in the biological context. While some filters (CR, IG) are broadly mentioned in the literature [18,16], some others, such as RL or CH, have been applied in the machine learning field but they are not so popular in genomic databases. Furthermore, we present here the results obtained by means of five criteria (filters CN, SP and wrappers WNB, WIB, WC4) that are barely used in this kind of data.

Table 1 reports the leave-one-out cross-validation (LOOCV) accuracy for each metric in Colon and Leukemia data sets. In the table, the first row shows

Table 1. LOOCV accuracy results for each classifier and gene selection technique

	Colon								Leukemia							
	NB (full 58.06)				C4 (full 80.65)				NB (full 100%)				C4 (full 73.61%)			
	(3)	(5)	(10)	(20)	(3)	(5)	(10)	(20)	(3)	(5)	(10)	(20)	(3)	(5)	(10)	(20)
SP	79.0^+	80.6^+	69.3	80.6^+	64.5	83.8	80.6	93.5^+	98.6	94.4	94.4	95.8	88.8^+	87.5^+	84.7	81.9
IG	85.4^+	85.4^+	85.4^+	80.6^+	85.4	74.1	85.4	85.4	94.4	93.0	94.4	95.8	90.2^+	87.5^+	86.1^+	81.9
RL	82.2^+	85.4^+	85.4^+	83.8^+	85.4	85.4	79.0	83.8	90.2	94.4	95.8	95.8	91.6^+	94.4^+	88.8^+	86.1^+
CH	85.4^+	85.4^+	87.1^+	88.7^+	85.4	85.4	85.4	83.8	98.6	97.2	95.8	97.2	88.8^+	84.7^+	83.3	81.9
CR	88.7^+	87.1^+	87.1^+	82.2^+	85.4	85.4	85.4	85.4	98.6	94.4	95.8	95.8	88.8^+	87.5^+	83.3	81.9
CN	85.4^+	85.4^+	87.1^+	87.1^+	85.4	85.4	83.8	85.4	98.6	97.2	95.8	97.2	88.8^+	88.8^+	83.3	81.9
WNB	82.2^+	87.1^+	85.4^+	87.1^+	85.4	80.6	69.3	74.1	95.8	95.8	95.8	95.8	88.8^+	91.6^+	83.3	81.9
WIB	62.9	67.7	77.4^+	79.0^+	82.2	77.4	77.4	88.7	98.6	95.8	94.4	97.2	88.8^+	88.8^+	86.1^+	84.7
WC4	88.7^+	85.4^+	85.4^+	85.4^+	85.4	83.8	83.8	85.4	94.4	95.8	95.8	95.8	88.8^+	91.6^+	83.3	81.9

the dataset and the second one the classifier next to the LOOCV percentage accuracies for non-gene selection for each classifier (in brackets). The rest of rows show the LOOCV values obtained by each method (first column) for each specified gene subset cardinality (3, 5, 10, 20). Furthermore, we conduct Studentt's paired two-tailed t-test in order to evaluate the statistical significance of the difference between the accuracy of each approach with gene selection and the result of the full set. Thus, the symbol " $+$ " and " $-$ " respectively identify statistically significant, at 0.05 level, wins or losses over the full set.

The top-ranked genes using the nine measures in each data set is listed next. All showed genes appear in the top-20-scoring lists of five ranking at least.

- **Colon:** R87126, M76378(1), M63391, M76378(2), J02854, M76378(3), X12671, M22382, T96873, M26383.
- **Leukemia:** X95735_at, M23197_at, M27891_at, U46499_at, M84526_at, L09209_s_at, D88422_at, M31523_at, M83652_s_at, M92287_at.

With regard to Colon domain, as we can see from table 1, for NB classifier, in all cases, except for SP(10) (i.e. subset with the top-10 genes from Soap list) and WIB(3)(5), these accuracy differences between the non-gene selection and the gene subset selected are statistically significant at 0.05 level. For C4 classifier, no statistical significant differences are shown between the accuracy of all the gene subsets selected by ranking metrics, except SP(20) for C4 classifier, and the accuracy of whole gene set. In some cases (most of then for C4 classifier), the classification accuracy is not improved when the number of genes of the subset is increased. In most of the cases, the accuracy obtained with the three first genes is the same or better than that obtained with the full set. Ranking provided by CR measure obtain the best averaged performance for the two classifier. An analysis of the genes selected by different approaches reveals interesting questions:

- Among the first 20 genes scored by the nine measures, the following two genes appear in the top-20-scoring lists of all scores (GenBank number): R87126, M76378(1).
- The following three genes appear eight times in the top-20: M63391, M76378(2), J02854.

- The following four genes appear seven times in the top-20: M76378(3), X12671, M22382, T96873.
- M63391 and M26383 (human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds.) appear seven and five times respectively in the top-3. The three clones of M76378 appear in the top-20 of seven rankings, and one version in two rankings.
- Most of the genes selected by evaluation measures appear in the lists of relevant genes detected by previous studies over this data set [9,8,2].
- ib and rl are measures of the same type (distance measures), but their lists are different. The same occurs with sp and cn, both of them are consistency measures and they have different rankings. However, ig and c4 using information measures and ch and cn have almost the same top-10.

Regarding Leukemia domain, for C4 classifier, when the cardinality of the subset is 3 or 5, the accuracy differences between the non-gene selection and the gene subset selected are statistically significant at 0.05 level for all cases. For NB classifier, no statistical significant differences are shown between the accuracy of all the gene subsets selected by ranking metrics and the accuracy of whole gene set. Also, as we can observe, most of the top-3 subsets obtain better results than the rest. There is not difference at the averaged performance of the nine ranked-lists. NB classifier obtains better results than C4, although such results are no statistically significant. An analysis of the genes selected by different approaches in Leukemia data set reveals the following interesting questions:

- Among the firs 20 genes scored by the nine measures, the following six genes appear in the top-20-scoring lists of all scores (GenBank number): X95735_at, M23197_at, M27891_at, U46499_at, M84526_at, L09209_s_at.
- The following six genes appear eight times in the top-20: D88422_at, M31523_at, M83652_s_at, M92287_at, X62320_at, M11722_at.
- X95735, M23197 and M27891 appear seven, six and seven times respectively in the top-3.
- Most of the genes selected by proposed evaluation measures appear in the lists of relevant genes detected by previous studies over this data set [9,8,2]. Note that these twelve genes are located almost at the same position in [2] with TNoM score.
- The all top-20 are very similar, emphasizing sp, ig, ch, cr and cn with 10 genes.

3.2 Classification with FSS Approaches

In this section, we empirically evaluate the efficiency and effectiveness of three subset evaluation measures (see Section 2.2) combined with a sequential forward search engine. LOOCV accuracy results for each gene selection algorithm are: 1) For Colon data set and NB classifier, 85.48^{+}, 85.48^{+} and 91.94^{+}, with correlation, consistency and wrapper subset evaluation measure respectively; and with C4, 88.71, 91.94 and $96,77^{+}$, respectively. 2) For Leukemia and NB, 98.61, 94.44 and 98.61 for the three measure respectively; and with C4, 81.94, 94.44^{+} and

94.44^{+}. With the aid of the wrapper gene selection technique, the two classifiers improve their results in the two data sets with respect to the ranking approach. In most of the cases, except in colon data set for NB classifier, accuracy differences between the wrapper procedure and the full set are statistically significant at 0.05 level. Besides, the consistency approach of the sequential search procedure wins over the full set in two cases, while correlation approach once. Results obtained with the wrapper approach are better than those obtained with the two filter techniques in all the cases except two on leukemia data set. This is due to the fact that subsets obtained by wrapper approaches will be better suited to the subsequent classification. The novelty of the application of wrapper approaches within biological data sets constitute a technique that has been proved to have a very good performance.

In both data sets, we notice the low number of genes selected by the consistency and wrapper approaches. In Colon domain, wrapper algorithm choose seven (H20709, M84326, H50623, M63391, H78386, R80427 and H23975) and six (R39465, H08156, J02854, D00860, R08021 and M26383) genes for NB and C4 classifiers respectively. We obtain two different subsets with wrapper approach because the process depend on the employed classifier, but only one subset with filter approaches, consistency five genes (M63391, D14812, T52015, K03460 and R87126), while correlation subset evaluation provide twenty–six genes. In Leukemia domain, wrapper choose three genes (D49950, D88422 and V68162) and two (M27891 and M195507) for NB and C4, three for consistency (M23197, AF009426 and AC002115) and fifty-one for correlation approach.

Sequential forward search procedure starts with an empty set and evaluates each gene individually to find the best single gene. It then tries each of the remaining genes in conjunction with the best to find the most suited pair of genes. In the next iteration each of the remaining genes are tried in conjunction with the best pair to find the most suited group of three genes. This process continues until no single gene addition improves the evaluation of the subset. Therefore, always choose the gene with the best individual evaluation, but generally the rest of the genes are not located at first positions of any ranked list of genes. Gene interactions can be captain for the subset selection approaches. All gene subset selection techniques are able to considerably reduce the huge number of genes to small informative and accurate subsets of components.

However, these accuracy improvements of wrapper procedures are couple with demanding computer-load necessities. In most of the cases, the computer-load necessities of ranking procedures can be considered as negligible with respect to wrapper ones. Consistency approach took 3 and 14 seconds to produce results on colon and leukemia domain respectively, correlation 26 and 1440 seconds, and wrapper took 165 and 1156 for NB classifier, and 520 and 309 seconds for C4.

4 Conclusions

Traditional feature selection methods often select the top-ranked features according to their individual discriminative power. When the number of features

is high, about thousands, as it happens in the microarray gene expression data sets, there are many irrelevant and/or redundant genes. For this reason, gene rankings might not be useful to select the best k genes from that ranked-list.

In this paper, we show that the classification accuracy may vary depending on the number of genes selected from the ranked–list, and not always is better when more genes are involved. In fact, it depends on the feature ranking method and also on the classifier. To show this situation, we have used nine feature ranking methods together with two different classifiers.

In addition, due to the effect of irrelevant and redundant genes in microarray gene expression data sets, those rankings might provide some noise to the classifier when we select the k top-ranked genes. This reason motivated us to study an algorithm to extract a subset of genes, trying to avoid the influence of unnecessary genes on the later classification. The wrapper approach of this algorithm shows an excellent performance, obtaining subsets better suited to the subsequent classification.

References

1. U. Alon et. al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745Ũ–50, 1999.
2. A. Ben-Dor et. al. Tissue classification with gene expression profiles. *Proc. Natl. Acad. Sci. USA*, 98(26):15149–54, 2001.
3. M. Dash, H. Liu, and H. Motoda. Consistency based feature selection. In *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pages 98–109, 2000.
4. T. Golub et. al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–37, 1999.
5. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
6. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machine. *Machine Learning*, 46(1-3):389–422, 2002.
7. M. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato, Dept Computer Science, Hamilton, New Zealand, 1999.
8. T. Hellem and I. Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome Biology*, 3(4):0017.1–0017.11, 2002.
9. I. Inza et. al. Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial Intelligence in Medicine*, 31:91–103, 2004.
10. I. Kononenko. Estimating attributes: Analysis and estensions of relief. In *European Conf. on Machine Learning*, pages 171–182, Vienna, 1994. Springer.
11. H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *7th IEEE Int. Conf. on Tools with Artificial Intelligence*, 1995.
12. H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. on Knowledge and Data Eng.*, 17(3):1–12, 2005.
13. J. R. Quinlan C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, California, 1993.
14. R. Ruiz, J. Riquelme, and J. Aguilar-Ruiz. Projection-based measure for efficient feature selection. *Journal of Intelligent and Fuzzy System*, 12(3–4):175–183, 2002.

15. I. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2005.
16. E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. In *Proc. 18th Int. Conf. on Machine Learning*, pages 601–608. Morgan Kaufmann, San Francisco, CA, 2001.
17. M. Xiong, L. Jin, W. Li, and E. Boerwinkle. Computatinal methods for gene expression-based tumor classification. *BioTechniques*, 29:1264–70, 2000.
18. L. Yu and H. Liu. Redundancy based feature selection for microarry data. In *10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2004.