ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Selection of more than one gene at a time for cancer prediction from gene expression data

Oleg Okun[1]*, Nikolay Zagoruiko[2]†, Alexessander Alves[3],
Olga Kutnenko[2] and Irina Borisova[2]

1- Inf Proc Lab - Dept of Electrical and Inf Eng
P.O.Box 4500, 90014 Oulu - FINLAND

2- Sobolev Inst of Mathematics - Russian Academy of Sciences
Koptyug avenue 4, 630090 Novosibirsk - RUSSIA

3- Lab of Artif Intell and Comp Science - University of Porto
Rua do Campo Alegre 823, 4150 Porto - PORTUGAL

Abstract. A new gene selection method capable of selecting more than one gene at a time is introduced. This characteristic contrasts it with almost all known methods assuming that there are no interactions between genes. The only exception is the pairwise gene selection method recently proposed by Bø and Jonassen [3]. Motivated by this method, we compare it and ours. Classification into healthy tissue and cancerous tumour is studied, where gene selection finds gene subsets well suitable for discriminating between these two classes.

1 Introduction

Microarrays are used to obtain expression levels for thousands of genes at once. They can greatly help in studies of different diseases at a molecular level and in design of new drugs preventing or curing these diseases. As a result of microarray experiments, gene expression matrices are produced. These matrices consist of many thousands of columns corresponding to genes whereas the number of rows, associated with the number of samples taken from patients, rarely exceeds a hundred. We consider the case when the samples are labelled (e.g. healthy/diseased). Thus, the task is to find sets of genes discriminating well between classes. To accomplish this task, various classifiers are employed in order to perform class prediction for new data by using the identified genes. Since gene expression data has many more attributes (features) than samples, many attributes can be safely removed since they are either noisy or irrelevant for class prediction. Hence, attribute selection is typically applied in order to find a small subset of the original attributes allowing good discrimination between experimental classes.

A vast majority of gene selection methods treats genes in isolation from each other, meaning that there are no interactions and interdependencies between

*Financial support from the Riitta and Jorma J. Takanen Foundation (Finland) is gratefully acknowledged.

†NZ, OK and IB work is supported by the grant 05-01-00241 from the Russian Foundation for Basic Research.

genes. However, the recent results in molecular biology prove the opposite. Based on them, Bø and Jonassen proposed a pairwise gene selection method in [3] and demonstrated on two public datasets that evaluating pairs of genes reveals what cannot be discovered if genes are analysed separately from each other.

Inspired by [3], a new gene selection method capable of selecting more than one gene at a time is introduced in this paper[1]. This characteristic contrasts it with all but one known methods (the only exception is [3]). However, whereas [3] can select genes in pairs, our method goes one step further; besides pairs, it can also select triples and higher order combinations of genes if desired. In addition, many methods require to pre-define the number of genes to be found, which implies one extra parameter to handle. Unlike the others, including [3] as well, our method automatically terminates when classification accuracy (acting as the objective function) achieved on the current subset begins to decrease. This happens because our method belongs to the wrapper model of feature selection while [3] is the filter model. In other words, unlike [3], our method carries out a search for a good subset using the induction algorithm itself as a part of evaluation, i.e. attribute selection is wrapped around the induction algorithm. Cross-validation (either leave-one-out or n-fold) is used to evaluate the current subset of attributes [1, 3, 4, 7, 8]. As was pointed in [7], "the disadvantage of the filter approach is that it totally ignores the effects of the selected feature subset on the performance of the induction algorithm". As a result, the filter model can find attributes that interact with the induction algorithm's bias counterproductively as remarked in [4].

2 Hillclimbing in attribute selection

Hillclimbing is one of the best known procedures for sequential attribute selection. Greedy algorithms such as backward elimination (BE) and forward selection (FS) implement so called unidirectional hillclimbing, i.e. attributes once added (removed) cannot be later deleted (added). FS starts with the empty subset and adds attributes one-by-one. At each step, the attribute yielding the best performance (e.g. lowest cross-validation error) of the current subset is added. BE starts with all attributes in the subset and removes them one at a time. At each step, the attribute whose removal leads to the best performance of the current subset is removed.

FS and BE algorithms can be improved by bidirectional hillclimbing when at each step the algorithm greedily adds n_1 attributes or deletes n_2 attributes as long as accuracy does not degrade. We will call bidirectional hillclimbing the AdDel algorithm. The advantage of AdDel compared to either FS or BE is that one or several previously deleted (added) attributes can be brought back to (removed from) the subset if the accuracy of the induction algorithm increases.

[1]Though there are attribute weighting methods like [6], weighting all attributes in parallel according to their relevance for classification, we do not consider them here since they typically do not take into account attribute interaction.

This advantage was stressed in several works [4, 8]. Another advantage of either hillclimbing procedure is that it can automatically determine the number of useful attributes, i.e. the user does not have to specify this parameter in advance in contrast to many filtering approaches.

3 GRAD algorithm

GRAD is an extension of AdDel applied to subsets of attributes instead of individual attributes. We call these subsets granules, hence the name 'GRAD', i.e. **GR**anular **A**d**D**el. GRAD is the wrapper method of feature selection involving two main steps.

- Form individual granules and the working set of granules.

  Step 1. Given n original attributes, rank individual attributes according to their prediction accuracy as measured by a classifier D.

  Step 2. Select top n_1 attributes from the ranked list. They form 1-granules.

  Step 3. Form all possible combinations of k out n_1 attributes, where $k = 2, \ldots, k_{max}$. Thus the number of combinations of attributes is equal to $\binom{n_1}{k}$ given k.

  Step 4. For each k, $k = 2, \ldots, k_{max}$, rank all $\binom{n_1}{k}$ combinations according to their prediction accuracy as measured by a classifier D.

  Step 5. For each k, $k = 2, \ldots, k_{max}$, select top n_k out $\binom{n_1}{k}$ combinations of attributes from the ranked list. They form 2-granules, ..., k_{max}-granules.

  Step 6. Form the working set consisting of $(\sum_{k=1}^{k_{max}} n_k)$ granules, each containing k attributes, where $k = 1, 2, \ldots, k_{max}$.

- Run AdDel on the working set of granules. A list of relevant granules is empty.

  Step 7. FS step: add (one-by-one) ℓ_1 best granules to the list of relevant granules based on prediction accuracy of a classifier D.

  Step 8. BE step: delete (one-by-one) ℓ_2 ($\ell_2 < \ell_1$) least relevant granules from the list of relevant granules based on prediction accuracy of a classifier D.

  Step 9. Measure prediction accuracy. If it decreased compared to the previous FS/BE steps, save the list of relevant granules and halt; otherwise go to Step 7.

In this work, D is the weighted k-nearest neighbour ($k = 1, 3, 5$). This choice was motivated by the fact that though it is known that this algorithm is sensitive to irrelevant attributes [1], it does not have parameters to tune, which is desirable if a large number of attribute subsets has to be evaluated as in our

case. Weights are associated with the original attributes. The ith weight is equal to the number of times the ith attribute appears in the list of relevant granules after halting GRAD. Hence, weights reflect importance or relevance of attributes: the higher weight, the more relevant attribute.

Given two n-dimensional patterns x and y, the similary between them is determined by the following formula:

$$\sqrt{\sum_{i=1}^{n} w_i(x_i - y_i)^2} = \sqrt{\sum_{i=1}^{n} (\sqrt{w_i}x_i - \sqrt{w_i}y_i)^2} \ ,$$

where w_i is the weight of the ith attribute.

Prediction accuracy is measured by means of leave-one-out cross-validation (LOOCV).

4 Method of Bø and Jonassen

This filtering method evaluates how well a pair of attributes in combination distinguishes two classes. First, each pair is evaluated by computing the projections of the training data on the diagonal linear discriminant (DLD) axis when using only two attributes constituting this pair. Then the two-sample t-statistic is computed on the projected points and assigned to a given pair as its score. Bø and Jonassen proposed two variants of attribute selection based on pair scores, called 'all pairs' (exhaustive search) and 'greedy pairs' (greedy search).

The all-pairs variant targets all possible pairs of attributes. First, all pairs are sorted in descending order of their score. After that, the pair with the highest pair score is selected and all other pairs containing either attribute included in this pair are removed from the sorted list of pairs. Then the next highest-scoring pair is found from the remaining pairs and all other pairs containing either attribute in this pair are removed from the list, and so on. This procedure terminates when the pre-specified number of attributes is reached.

Since the all-pairs variant is computationally demanding, an alternative evaluating only a subset of all pairs is proposed (greedy pairs). The greedy-pairs variant first ranks all attributes based on indivudual t-score. Next, the best attribute a_i ranked by its t-score is selected. Among all other attributes, the attribute a_j that together with a_i maximises the pait t-score is found so that a_i and a_j form a pair. These two attributes are then removed from the attribute set and a search for the next pair is performed until the desired number of attributes is selected.

5 Experiments

We chose the dataset of the expressions of 822 genes in 74 samples. Its description can be found in [5]. The distribution of normal and cancerous samples is imbalanced with the bias toward the latter (24 samples are normal while 50 samples are cancerous).

For Bø and Jonassen's method, the number of attributes to be selected was set to 2, 4, 6, ..., 100.

For GRAD, we used $k_{max} = 3$, $\ell_1 = 10$, $\ell_2 = 5$, $n_1 = n_2 = n_3 = 50$. Settings for ℓ_1 and ℓ_2 reflect the fact that the number of selected genes should not be too small in order to achieve high generalisation accuracy. Though bidirectional hill-climbing can better cope with local optima than its unidirectional counterparts, the former cannot, however, totally avoid them. This implies that GRAD can typically halt after few FS/BE steps as was usually observed in our experiments. This effect prevented GRAD to select too many attributes, which is desirable from the biological point of view. To validate feature selection, 5-nearest neighbour was used.

LOOCV was used with both gene selection and classification. First, gene selection was done for each cross-validation (CV) subset, where LOOCV error was utilised in order to assess performance. Then genes selected for all CV subsets were gathered into a single histogram having n entries, equal to the number of genes. Using this histogram and a user-specified threshold T, rare genes were filtered out, because these genes are possibly selected due to random (noisy) factors[2]. The set of the remaining genes was then used to validate a classifier so that the final output is LOOCV classification error. A different classifier than the one used for CV of feature selection can be employed for CV of classification. We found that sometimes it is beneficial to LOOCV error.

Table 1 shows LOOCV errors when no gene selection is applied (2nd column) as well as when no rare gene filtering is done with GRAD (3th column) and when this filtering is used (4th column). Independently of whether filtering is used or not, using GRAD led to much lower errors than without it. Also rare gene filtering can result in lower error compared to the 'no filtering' case and this error is not necessarily achieved with the same classifier as the one used for CV of feature selection.

kNN	No gene selection	GRAD (no filtering)	GRAD ($T = 50$)
1NN	27.0	**16.2**	20.3
3NN	40.5	25.7	21.6
5NN	31.1	21.6	**16.2**

Table 1: LOOCV classification errors (in %) when no gene selection is done prior to classification and for GRAD with and without filtering out rare genes before cross-validating a classifier. The 1st column indicates an induction algorithm used to validate classification. Best results are shown in bold.

Without rare gene filtering, GRAD selected 107 genes in all granules of 74 CV subsets (individual gene occurrence varies from 1 to 121). The number of 1-granules, 2-granules, and 3-granules was 1,110 (80.43%), 162 (11.74%), and 108 (7.83%), respectively. When rare gene filtering ($T = 50$) was applied, indices[3] of

[2]Rare genes were the key reason why we did not test a classifier on the left-out samples for each CV subset, but instead we preferred to combine all selected genes together.

[3]We assume that the indices start at 1.

11 selected genes were: 48, 53, 155, 249, 333, 348, 372, 409, 437, 762, and 769.

Table 2 summarises results for Bø and Jonassen's method. AP and GP stand for all-pairs and greedy-pairs variants, respectively. LOOCV classification error was estimated with and without rare gene filtering. The 1st figure in each bracket indicates the number of genes to be selected, corresponding to the lowest LOOCV error for a given classifier and the variant of Bø and Jonassen's method. The 2nd figure in each bracket is the total number of selected genes accumulated from all 74 CV subsets (these genes were used to validate a classifier) for the number of genes to be selected specified by the 1st figure in brackets.

kNN	AP (no filtering)	GP (no filtering)	AP ($T = 50$)	GP ($T = 50$)
1NN	20.3 (10,40)	18.9 (90,200)	21.6 (16,12)	18.9 (86,77)
3NN	**16.2** (4,21)	24.3 (22,77)	**16.2** (24,20)	21.6 (32,26)
5NN	18.9 (4,21)	25.7 (12,56)	18.9 (24,20)	21.6 (4,3)

Table 2: LOOCV classification errors (in %) for Bø and Jonassen's method. The 1st column indicates an induction algorithm used to validate classification. Best results are shown in bold.

Though it seems that both GRAD and Bø and Jonassen's method demonstrated the same best performance (16.2% of LOOCV error), GRAD did so with 11 genes, while its competitor needed almost two times more genes (20). However, being a wrapper, GRAD is slower than Bø and Jonassen's method.

References

[1] D.W. Aha, Generalizing from case studies: a case study. In D.H. Sleeman and P. Edwards, editors, *proceedings of the* 9^{th} *International Conference on Machine Learning* (ICML 1992), pages 1-10, July 1-3, Aberdeen, Scotland, UK, 1992.

[2] H. Almuallim and T.G. Dietterich, Learning with many irrelevant features, In *proceedings of the* 9^{th} *National Conference on Artificial Intelligence* (AAAI 1991), pages 547-552, July 14-19, Anaheim, CA, 1991.

[3] T.H. Bø and I. Jonassen, New feature selection procedures for classification of expression profiles, *Genome Biology*, 3(4):research0017.1-0017.11, 2002.

[4] R. Caruana and D. Freitag, Greedy attribute selection. In W.W. Cohen and H. Hirsh, editors, *proceedings of the* 11^{th} *International Conference on Machine Learning* (ICML 1994), pages 28-36, July 10-13, New Brunswick, NJ, 1994.

[5] O. Gandrillon, Guide to the gene expression data. In *proceedings of the ECML/PKDD Discovery Challenge Workshop*, pages 116-120, Pisa, Italy, 2004.

[6] B. Hammer and T. Villmann, Generalized relevance learning vector quantization, *Neural Networks*, 15(8-9):1059-1068, 2002.

[7] G.H. John, R. Kohavi and K. Pfleger, Irrelevant features and the subset selection problem. In W.W. Cohen and H. Hirsh, editors, *proceedings of the* 11^{th} *International Conference on Machine Learning* (ICML 1994), pages 121-129, July 10-13, New Brunswick, NJ, 1994.

[8] P. Langley and S. Sage, Oblivious decision trees and abstract cases. In D.W. Aha, editor, *proceedings of the 1994 AAAI Workshop on Case-Based Reasoning* (AAAI 1994), pages 113-117, August 1-2, Seattle, WA, 1994.