Monte Carlo feature selection for supervised classification

Article *in* Bioinformatics · February 2008

DOI: 10.1093/bioinformatics/btm486 · Source: PubMed

6 authors, including:

Michał Dramiński
Polish Academy of Sciences
34 PUBLICATIONS **223** CITATIONS
SEE PROFILE

Alvaro Rada-Iglesias
Universidad de Cantabria
57 PUBLICATIONS **6,709** CITATIONS
SEE PROFILE

Stefan Enroth
Uppsala University
189 PUBLICATIONS **7,122** CITATIONS
SEE PROFILE

Claes Wadelius
Uppsala University
189 PUBLICATIONS **9,730** CITATIONS
SEE PROFILE

BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 1 2008, pages 110–117
doi:10.1093/bioinformatics/btm486

Data and text mining

Monte Carlo feature selection for supervised classification

Michał Draminski[1], Alvaro Rada-Iglesias[2], Stefan Enroth[3], Claes Wadelius[2], Jacek Koronacki[1,†] and Jan Komorowski[3,4,*,†]

[1]Institute of Computer Science, Polish Academy of Science, Ordona 21, PL-01-237 Warsaw, Poland, [2]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, [3]The Linnaeus Centre for Bioinformatics, Uppsala University and The Swedish University for Agricultural Sciences, Box 758, SE-751 24 Uppsala, Sweden and [4]Interdisciplinary Centre for Mathematical and Computer Modelling, Warsaw University, Poland

Received on December 13, 2006; revised on August 28, 2007; accepted on September 25, 2007

Advance Access publication November 28, 2007

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Pre-selection of informative features for supervised classification is a crucial, albeit delicate, task. It is desirable that feature selection provides the features that contribute most to the classification task *per se* and which should therefore be used by any classifier later used to produce classification rules. In this article, a conceptually simple but computer-intensive approach to this task is proposed. The reliability of the approach rests on multiple construction of a tree classifier for many training sets randomly chosen from the original sample set, where samples in each training set consist of only a fraction of all of the observed features.

Results: The resulting ranking of features may then be used to advantage for classification via a classifier of any type. The approach was validated using Golub *et al*. leukemia data and the Alizadeh *et al*. lymphoma data. Not surprisingly, we obtained a significantly different list of genes. Biological interpretation of the genes selected by our method showed that several of them are involved in precursors to different types of leukemia and lymphoma rather than being genes that are common to several forms of cancers, which is the case for the other methods.

Availability: Prototype available upon request.

Contact: jan.komorowski@lcb.uu.se

1 INTRODUCTION

A major challenge in the analysis of many biological data matrices is due to their sizes: a very small number of records (samples), of the order of tens, versus thousands of attributes or features for each record. This challenge is aggravated by noise, which is inherent to the biological data. An obvious example are microarray gene expression experiments (here, the features are genes or, more precisely, their expression levels). In such tasks, supervised classification is quite different from a typical data mining problem, in which every class has a large number of examples. In the latter context, the main task is to propose a classifier of the highest possible quality of classification. On the other hand, e.g. in class prediction for typical gene expression data, it is not a classifier *per se* that is crucial; rather, selection of informative genes and a reliable assessment of classification results is the most important issue. Given such data, all reasonable classifiers can be claimed to be capable of providing essentially similar results [if measured by error rate or the like criteria; c.f. Dudoit and Fridlyand (2003)].

As Dudoit and Fridlyand argue convincingly and in detail (Dudoit and Fridlyand, 2003), *The importance of taking feature selection into account when assessing the performance of a classifier cannot be stressed enough*. In particular, if the assessment procedure is based on cross-validation, it is beyond question that selection of informative genes should be done within the cross-validation procedure (c.f. Dudoit and Fridlyand, 2003), the references cited there and Simon *et al*. (2003).

The problem with this approach is that, while it provides as honest assessment of classifier's performance as possible, it makes feature selection a part of that classifier's training. However, one would like to have groups of genes that contribute most to the classification task, and hence are informative or 'relatively important', to a given classification task *regardless* of the classifier that is used. In other words, it is preferred to have an objective measure of relative importance of the genes for a particular classification task. One should also note that this issue is different from a general problem of determining differentially expressed features, and in particular, differentially expressed genes. For excellent accounts of this and other statistical issues in microarray data analysis see Speed (2003) and Smyth *et al*. (2003). In such a general setup, evidence for differential expression is ranked by statistical means and significance analysis of the results obtained is performed, but all this is done without a reference to any classification task.

In this article, we propose a procedure for selecting features to be included in the data for a given supervised classification task. The training set thus constructed can then be used to train any classifier. Our approach is fundamentally different from that of Su *et al*. (2003) that, although sound and interesting as it is, cannot take into account the fact that a feature may prove informative only in conjunction with some other features, but not alone.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint First Authors.

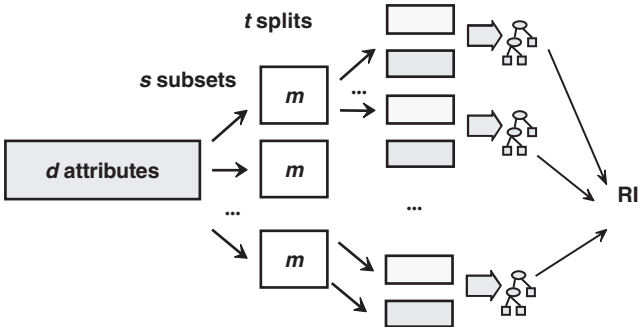

Fig. 1. Block diagram of the main step of the procedure.

The procedure is introduced in Section 2, and its use is illustrated on two data sets, namely the leukemia data of Golub *et al*. (1999) and the lymphoma data of Alizadeh *et al*. (2000), in Section 3. We conclude in Section 4 with a few remarks.

The Golub *et al*. (1999) data set comprises 38 samples from two classe: 27 cases of acute lymhoblastic leukemia (ALL) and 11 cases of acute myeloid leukemia (AML). Each case is described by expression levels of 7129 genes. The second data set consists of measurements of 4026 genes from 62 patients. The patients are classified into three classes: lymphoma and leukemia (DLCL; 42 samples), follicular lymphoma (FL; 9 samples) and chronic lymphocytic leukemia (CLL; 11 samples).

2 FEATURE SELECTION

2.1 The main step of the procedure

Our procedure is conceptually very simple, albeit computer-intensive. We consider a particular feature to be important, or informative, if it is likely to take part in the process of classifying samples into classes 'more often than not'. This 'readiness' of a feature to take part in the classification process, termed relative importance of a feature, is measured via intensive use of classification trees. The use of classification trees is motivated by the fact that they can be considered to be the most flexible classifiers within the family of all classification methods. So, the classifiers are used for measuring relative importance of features, not for classification as such.

In the main step of the procedure, we estimate relative importance of features by constructing thousands of trees for randomly selected subsets of features. More precisely, out of all d features, s subsets of m features are selected, m being fixed and $m<<d$, and for each subset of features, t trees are constructed and their performance is assessed. Each of the t trees in the inner loop is trained and evaluated on a different, randomly selected training and test sets that come from a split of the full set of training data into two subsets: each time, out of all n samples, about 66% of samples are drawn at random for training (in such a way as to preserve proportions of classes from the full set of training data) and the remaining samples are used for testing. See Figure 1 for a block diagram of the procedure.

In sum, in the main step of the procedure, st trees are constructed and evaluated. Both s and t should be sufficiently large, so that each feature has a chance to appear in many different subsets of features and that randomness due to inherent variability in the data is properly accounted for.

A crude measure of relative importance of a particular feature could be given as the overall number of splits made on that feature in all nodes of all st trees. Clearly, however, for any particular split, its contribution to the overall relative importance of the feature should be weighted by the information gain achieved by the split, the number of samples in the split node as well as by the classification ability of the whole tree.

In order to determine relative importance, let us first introduce weighted accuracy of a tree as a means to assess classification ability of the tree on a test set. Unlike un-weighted accuracy, Acc, weighted accuracy, to be denoted wAcc, takes into account sizes of the classes in such a way as to prevent undue influence of a majority class on the performance index. For a classification problem with c classes, let n_{ij} denote the number of samples from class i classified as those from class j; clearly, $i,j,=1,2,\ldots,c$ and $\sum_{i,j} n_{ij}=n$, the number of all samples. Now, one can define weighted accuracy as

$$wAcc=\frac{1}{c}\sum_{i=1}^{c}\frac{n_{ii}}{n_{i1}+n_{i2}+\cdots+n_{ic}}, \qquad (1)$$

i.e. as the mean of c true positive rates.

Further, if a particular split is made on feature g_k, then the more informative this feature is, the greater is wAcc for the whole tree. Also, the greater are the information gain on the split and the number of samples in the split node. Information gain can be measured, e.g. by Gini Index or Gain Ratio, and so the relative importance of feature g_k, RI_{g_k}, can be defined as

$$\mathrm{RI}_{g_k}=\sum_{\tau=1}^{st}(\mathrm{wAcc})^u\sum_{n_{g_k}(\tau)}\mathrm{IG}(n_{g_k}(\tau))\left(\frac{\mathrm{no.\,in}\,n_{g_k}(\tau)}{\mathrm{no.\,in}\,\tau}\right)^v, \qquad (2)$$

where summation is over all st trees and, within each τ-th tree, over all nodes $n_{g_k}(\tau)$ of that tree on which the split is made on feature g_k, $\mathrm{IG}(n_{g_k}(\tau))$ stands for information gain for node $n_{g_k}(\tau)$, $(\mathrm{no.in}\,n_{g_k}(\tau))$ denotes the number of samples in node $n_{g_k}(\tau)$, (no. in τ) denotes the number of samples in the root of the τ-th tree and u and v are fixed positive reals. Note that by taking, say, $u=2$ trees with low wAcc are penalized more severely than when taking $u=1$. Similarly, the greater is v, the smaller is the influence of node $n_{g_k}(\tau)$ with a given ratio $(\mathrm{no.in}\,n_{g_k}(\tau))/(\mathrm{no.in}\,\tau)$ on RI_{g_k}, unless $n_{g_k}(\tau)$ is the tree's root. And, for any fixed positive v, the influence of any particular node on RI_{g_k} decreases monotonically with the number of samples in this node. In this way and especially for low level nodes in a tree, the fact that information gains can be very high, while only very small subsets of data are split, is taken into account.

We also note that setting $u=v=0$ and taking $\mathrm{IG}(n_{g_k}(\tau))$ to power 0, we obtain a crude measure of the relative importance of feature g_k, to be referred to as cRI_{g_k}, equal to the overall number of splits made on g_k in all nodes of all st trees.

In the procedure, there are five parameters, m, s, t, u and v, to be set by an experimenter. A reasonable value of t, given m, can be provided relatively easily, so we shall consider it fixed. Moreover, it is not difficult to provide not one but several

rankings, each based on RI_{g_k} with a different pair of u and v values, say, with $u=0$, 1, 2 and $v=0.5$, 1, 2. The obtained rankings may be then compared and the issue of setting the value for the two parameters may be considered resolved as well.

The choice of subset size m of features selected for each series of t experiments should take into account the trade-off between the need to prevent informative features from being masked too severely by the relatively most important ones and the natural requirement that s be not too large. Indeed, the smaller the m, the smaller a chance of masking the occurrence of a feature. However, a larger s is needed, since all features must have a high chance of being selected into many subsets of the features. (We shall comment more on the issue of masking in the sequel.)

We suggest performing the ranking procedure several times—each time with another value of m. In our scrutiny of the leukemia and lymphoma data, we experimented with $m=50$, 100 and 300 and obtained essentially the same results. Now, for a given m, s is made a running parameter of the procedure, and the procedure is executed for $s=s_1$, s_1+10, $s_1+20, \ldots$ until the rankings for successive values of s [and each fixed (u,v)-pair] of top p% features prove (almost) the same. Minimal number of subsets, s_1, is in fact random and is such that the ranking based on these subsets includes p% of all features in the full data sample. Note that after having used s subsets of m features, at most sm features can be ranked, and the probability of achieving this upper bound is practically zero. More precisely, a distance between two successive rankings is defined, and the procedure is run until the values of the distance stabilize at some acceptably low level, i.e. close to zero, for all (u, v)-pairs used in Equation (2). Note that for each s, several rankings are provided, each for a different (u, v)-pair. The distance between the ranking obtained after s subsets of m features have been used in the procedure and the ranking reached after using $s-10$ subsets is defined as follows:

$$\mathrm{Dist}(s, s-10) = \frac{1}{d_p} \sum_{g_k} |\mathrm{rank}(g_k, s) - \mathrm{rank}(g_k, s-10)| \qquad (3)$$

where summation is over top p% features obtained after having used $s-10$ subsets. $\mathrm{Rank}(g_k, r)$ is the rank of feature g_k after having used r subsets, and d_p is the normalizing constant equal to the number of features taken into account ($d_p = \mathrm{dp}/100$). Parameter p should not be too large and it is suggested that it lies between 5 and 20. According to our experience, any value from this range may be chosen since essentially the same results are attained. However, the value of p should not be too large, say, greater than 40. Indeed, since only a small fraction of all features is truly informative, then, for large p, features from the bottom of one ranking can change ranks very dramatically in the successive ranking, thus making the distance unduly influenced by such artifact.

2.2 Validation and confirmatory steps

Given the nature of the data we are interested in, namely their exceedingly large dimension (number of features) compared to the number of samples, special emphasis must be placed on checking statistical significance of the results obtained. We propose to incorporate into the whole procedure two validation steps and an additional confirmatory step.

In brief, for a given data set, the first validation step consists in repeating the main step of the procedure with, say, 50 different permutations of class labels of the samples. The aim is to show that the classification results obtained earlier measured by the distribution of wAcc on all st trees are significantly different from what can be obtained under randomly permuting the labels (classes) of samples, hence making the class independent of feature values. It is thus a way to confirm that the data are informative, i.e. that there exists some evident connection between feature values and classes. This fact justifies the search for the most important features, based on the data provided.

The second validation step consists in showing that reliable class prediction can be performed using only a few, say b, randomly chosen features out of $2b$ features earlier found to be relatively most important. This is done by again constructing thousands of trees on b features out of the $2b$ most important features, as well as on randomly selected sets of b features from the set of the remaining $d-2b$ features. For each set of b features, many training and testing sets of samples are drawn at random from the original set of samples and trees are trained and tested on these sets. Roughly speaking, in this step, trees are constructed on samples with just b features, either b features from the $2b$ most important ones or b features chosen at random from all the remaining ones. As a result, two distributions of wAcc are obtained, one for b features from the $2b$ most important ones and another for b features chosen at random from all the remaining ones, with the goal to prove significance of classification results for the best features.

The idea behind this last validation step is clear. The distribution of weighted accuracy of trees for randomly chosen vectors of features serves as a fiducial distribution under the hypothesis of using non-informative features. The wAcc distribution for vectors with features claimed most important is then placed against the fiducial distribution to validate the claim. If successful, and given in particular that the classification abilities of all trees are measured on test samples, such a validation is conclusive for the whole data set.

At the same time, however, our validation does not provide anything like an 'exact level of significance' of the results obtained. Extensive resampling introduces intrinsic interdependencies within the whole procedure, making results conditional on the data. Hence, we propose one additional confirmatory step in the procedure. It consists in splitting first the data set into two subsets, comprised of ~75% and 25% of the whole data, respectively. The first subset is termed the final validation set and the latter—the final test set. Then, the main step of the procedure is run on the final validation set, and the earlier described second validation step is run with wAcc's calculated on the basis of the final test set (not used in the main step in any way). Hence, the obtained statistical significance of the results on the most important features may be claimed unconditionally true. It is a different matter that it pertains to data sets consisting of fewer samples than the original data set. In any case, the claim that the results obtained earlier (based on all samples) are unconditionally significant gets ground too.

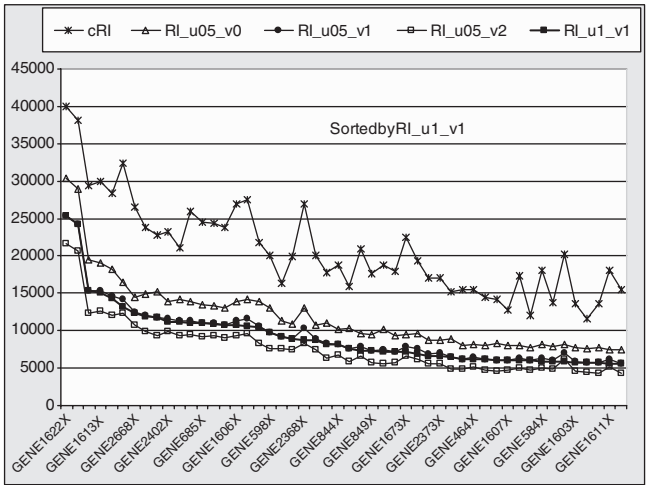

Fig. 2. Rankings for different (u, v)-pairs: lymphoma data.

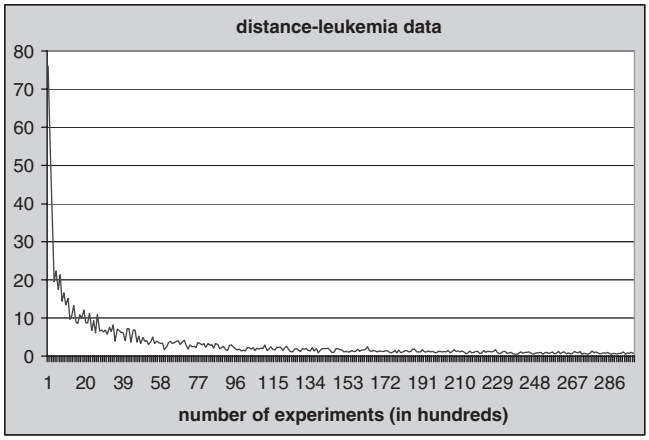

Fig. 3. Distance between the rankings (for top 5% features) as a function of s: leukemia data.

3 EXPERIMENTS AND RESULTS

Two experiments on well-known data sets were performed. First, we describe the experiments and then compare quantitatively our results with the results given by other approaches. And second, we proceed with biological analysis of the ranked genes by examining literature and their GeneOntology annotations (The Gene Ontology Consortium, 2000).

3.1 Gene selection

In all the experiments, C4.5 trees (version 8) were constructed, as implemented in WEKA 3-4-1 (Wittenn and Frank, 2005) (actually, j48 tree was used with the same parameters throughout the whole study, in particular with ConfidenceFactor = 0.25 for pruning). Trees were grown on the original data of Golub *et al.* (1999). That is, no preprocessing of the data was performed unlike, e.g. in the work of Dudoit *et al.* (2002, 2003). In the case of the data of Alizadeh *et al.* (2000), trees were grown separately on the original data, with missing values left intact, and on the data with missing values imputed in the same way as it was done in Tibshirani *et al.* (2003).

The parameters of the main step of the procedure were chosen following the suggestions given above. In particular, relative importance for both data sets [see Equation (2)] was measured for several pairs of parameters u and v. Also, cRI_{g_k}, the crude measure of relative importance, was used to build yet another ranking of genes. Somewhat surprisingly, the rankings obtained for different (u, v)-pairs have proven very similar. Sample of results are presented in Figure 2, where the rankings for the lymphoma data of Alizadeh *et al.* are given for cRI_{g_k} and RI_{g_k} with the following (u, v)-pairs: (0.5,0), (0.5,1), (0.5,2) and (1,1), and the genes are ordered according to RI_{g_k} with $u = v = 1$.

Analysis of the distance function given by Equation (3) determined the number of subsets of features s required for obtaining a stable ranking. Sample analysis for the leukemia data of Golub *et al.* with $m = 300$, $t = 10$ and $u = v = 1$, is presented in Figure 3. Each of the trees in the inner loop (i.e. each of the 10 trees) was trained and evaluated on different, randomly selected training and test samples. Each time 66% of the samples were drawn at random for training in such a way as to preserve proportions of the classes in the original set of data with the remaining samples used for testing. The same analysis was performed for the lymphoma data, leading to essentially the same picture.

It follows from Figure 3 that in order to obtain stable rankings for each of the data sets, other parameters being fixed, it more than suffices to choose $s = 3000$, i.e. to select 3000 subsets of 300 genes (out of 7129 genes in the case of the Golub *et al.* data and, by a similar analysis, out of 4026 genes in the case of the Alizadeh *et al.* data).

For the leukemia data of Golub *et al.* the 30 relatively most important genes that we have obtained are as follows (genes are given in columns, from the 1st till the 30th):

1. X95735_at	11. M54995_at	21. M16038_at
2. M31166_at	12. U02020_at	22. D14874_at
3. M27891_at	13. M77142_at	23. M84526_at
4. M55150_at	14. M81933_at	24. M92287_at
5. D88422_at	15. X70297_at	25. M31523_at
6. M23197_at	16. Y12670_at	26. X62654_rna 1_at
7. M98399_s_at	17. U46499_at	27. M31303_rna 1_at
8. U50136_rna 1_at	18. M83652_s_at	28. D49950_at
9. M21551_rna 1_at	19. U46751_at	29. U22376_cds2_s_at
10. M27783_s_at	20. L09209_s_at	30. J05243_at

It is worthwhile to compare genes selected by our method versus genes chosen by Dudoit *et al.* (2002, 2003), where over-expressed genes were found using multiple hypothesis testing. (We notice that there is a significant overlap between the list of Dudoit *et al.* and that of Golub *et al.*) In Dudoit *et al.* (2002), 32 over-expressed genes were found for the AML cases and 60 over-expressed genes for the ALL cases. One should expect that over-expression of a gene is not necessary for its importance for classification and hence there should not be too much overlap between the set of high ranking genes and that of over-expressed ones. For example, among the 30 top ranking genes selected by our procedure, there are 12 that are not

over-expressed and hence are not on the list provided by Dudoit *et al*. In the table below, we give a comparison of classification ability (measured by Acc and wAcc) of these 12 non-over-expressed genes against the 12 top ranking genes also selected by our procedure but that are on the Dudoit *et al*. list. The two thus obtained sets of samples, each sample comprising 12 features, were used to train the following classifiers: C4.5 (J48), 1 Nearest Neighbor, Naive Bayes, Random Forest and Support Vector Machine (all as implemented in WEKA 3-4-1, with default parameters). For each set of samples and each classifier, 10-fold cross-validation was performed 100 times (via randomly reordering the samples) and the mean Acc and wAcc (the latter given in brackets) were calculated:

	J48	1NN	NB	RF	SVM
non-over-expressed	0.918 (0.901)	0.972 (951)	0.997 (0.998)	0.950 (916)	0.955 (0.923)
over-expressed	0.937 (0.949)	0.947 (0.936)	0.974 (0.981)	0.958 (0.934)	0.947 (0.935)

No doubt, over-expression is indeed not needed for a gene to contribute highly to classification. Moreover, our method is capable of exploiting interactions between features and hence finding groups of features that together contribute to classification. A separate study on discovering such instances is now pending.

For the ranking of the imputed lymphoma data of Alizadeh *et al*. the 30 relatively most important genes (in GID notation) found by us are:

1. GENE1622X
2. GENE1602X
3. GENE1616X
4. GENE654X
5. GENE1619X
6. GENE2402X
7. GENE1702X
8. GENE2400X
9. GENE1617X
10. GENE1647X
11. GENE2403X
12. GENE1553X
13. GENE1744X
14. GENE530X
15. GENE1613X
16. GENE1618X
17. GENE1610X
18. GENE712X
19. GENE1644X
20. GENE2399X
21. GENE622X
22. GENE653X
23. GENE1662X
24. GENE625X
25. GENE833X
26. GENE1606X
27. GENE620X
28. GENE588X
29. GENE1643X
30. GENE598X

For the data of Alizadeh *et al*. it is worthwhile to compare the ranking of genes obtained by our method with the set of genes found important by the nearest shrunken centroids method of Tibshirani *et al*. (2003). In that latter study, 81 genes were found important for classification. It appears that our list of the 81 relatively most important genes includes only 29 genes from the 81 genes found important by the method of Tibshirani *et al*.; among our 300 most important genes there are 48 from their list of 81 genes. Their study was based on 59 samples (3 outlying samples from the whole set of 62 samples were disregarded). In order to evaluate the method, Tibshirani *et al*. selected at random

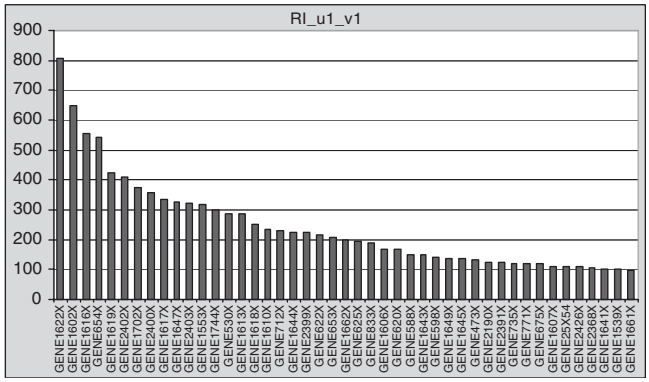

Fig. 4. Relative importance for the 45 most important genes: lymphoma data.

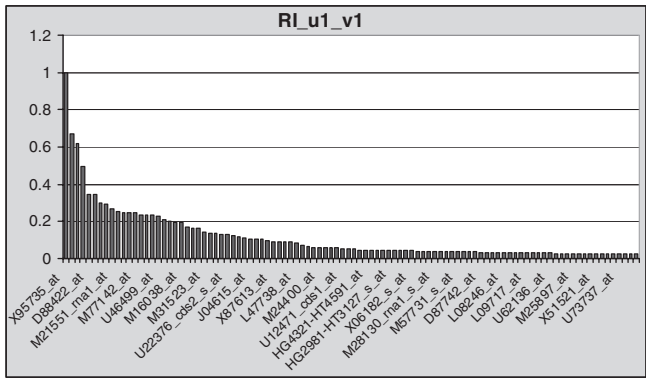

Fig. 5. Relative importance for the 100 most important genes: leukemia data.

20 samples as a test set and, for their set of 81 genes, trained their classifier on the remaining 39 samples to find that it made no classification error. For comparison, we built 30 000 trees on randomly selected 30 000 training sets of 42 samples and assessed them on the corresponding test sets of 20 samples, where each tree was built on just 45 features chosen at random from the 90 features found most important by our procedure. The results proved very encouraging, with some 8% cases with zero errors, about 75% cases with at most 3 errors and a few cases with the maximum number of 7 errors.

The relative importance of genes is given in Figure 4 for the imputed lymphoma and in Figure 5 for the leukemia data. It is seen that, except for a few most important genes, the rankings are rather flat.

We omit here a detailed discussion of the corresponding results for the Alizadeh *et al*. data with missing values left intact, since the general conclusions are rather similar. However, the obtained ranking differs rather markedly from that for the imputed data. In the two rankings of 100 most important genes, only 60 are included in both and the intersection of the two sets of the 200 most important genes contains 123 genes.

The rankings should be checked for their possible susceptibility to, or hopefully robustness against, the effect of masking some features by co-related features that happen to be more important in terms of their capacity to reduce diversity of classes in children nodes of a tree. An obvious way to prevent this effect from occurring is to include the so-called surrogate splits into our measure of importance. See Breiman *et al.* (1984) for the definition and discussion of those splits. Nevertheless, we have decided not to do so, since we expected only a few of all the features to be truly important. Hence, only features found as the highest ranking ones deserve further scrutiny whether they mask any other features or not. To put it otherwise, only for the most important features finding surrogate splits might be of interest.

Accordingly, we suggest proceeding in two stages: first, to rank genes according to the RI_{g_k} measure and then to check whether the obtained most important genes obscure the ranking of the remaining genes in any way. Our way to perform the latter stage is to remove a few, say 5–10, highest ranking genes from all the samples, repeat the main step of the procedure and see if the ranking of the remaining genes has changed in any (significant) way. Of course, removing top few genes and repeating the main step of the procedure can be iterated several times. For our two example data sets, such an analysis did not lead to any significant changes in the rankings. In both examples, we iterated the removal of the 6 top ranking genes 10 times and found that the changes were minor, if any. This particularly was the case for the data of Golub *et al.*

In the case of the leukemia data of Golub *et al.* it can readily be seen that gene X95735_at is the only gene that is capable of discerning between AML and ALL samples in just one split. This can be implemented by constructing a tree and looking for surrogate splits. The situation is different for the lymphoma data, both imputed and left intact. However, we shall confine ourselves to describing the former case only. In this case there are 7 genes, (1622X, 1602X, 1616X, 1619X, 1702X, 1617X and 1618X ranked, respectively, 1st, 2nd, 3rd, 5th, 7th, 9th and 16th), which require just 1 split on any of them to discern DLCL samples from those from the other two classes. Moreover, one split on gene 1671X (ranked 76th) separates FL samples from the rest, and there are 28 genes on which CLL samples can be separated from the rest by 1 split, too. Twelve of them have ranks below 100, 19 below 200, 25 below 300 and only 1 has rank below 400, namely 492. It follows from the above that masking is not a serious problem, at least not in the case of the examples that we investigated.

Statistical correctness of the procedure was assessed by performing the validation and confirmatory steps, as described in Section 2.2, to the effect that the rankings obtained may safely be claimed significant. Due to space limitations, this assessment is accessible at http://www.ipipan.eu/staff/m.draminski/files/supplement1203.pdf.

At the same place, the reader may find a full justification of the claim that, regardless of a classifier to be used, the features that rank best according to our procedure indeed do contribute more than the other features to the classification problem under scrutiny. Here, we only give a summary of the justification of the claim.

The rationale behind the choice of classification trees as the building blocks from which the ranking follows is that trees are flexible classifiers since disjunctions of conjunctions (and this is how class assignment is provided) can model arbitrarily complex decision surfaces. Of course, the problem that no classifier, even if flexible in principle, is perfect, remains. To this end, we have repeated the whole procedure with each tree replaced by a totally different rule-based classifier. For our two example data sets, we have found that the groups of top ranking genes obtained by the two procedures have a sufficient overlap to support the claim.

More importantly, in lieu of C4.5 (J48) we performed the second validation step from Section 2.2 using the following classifiers (with default parameters): 1NN, NB, RF and SVM. For the Alizadeh *et al.* data either 45 out of 90 best genes or 45 out of the remaining 3936 genes were used for classification; for the Golub *et al.* data, 100 out of 200 best genes or 100 out of the remaining 6930 genes were used. In each of the 4 cases, 100 sets of genes were drawn at random. In the following table, median *Acc* and *wAcc* (the latter in brackets) for the 100 experiments are given, confirming the validity of our claim:

	Alizadeh *et al.*		Golub *et al.*	
	45 out of 90	45 out of 3936	100 out of 200	100 out of 6930
J48	0.93 (0.89)	0.78 (0.78)	0.86 (0.82)	0.71 (0.67)
1NN	1.0 (1.0)	0.93 (0.97)	0.93 (0.95)	0.78 (0.76)
NB	1.0 (1.0)	0.71 (0.90)	0.93 (0.90)	0.86 (0.92)
RF	0.93 (0.97)	0.86 (0.92)	0.93 (0.92)	0.78 (0.88)
SVM	1.0 (1.0)	0.93 (0.97)	0.93 (0.95)	0.86 (0.82)

3.2 Biological validation

The validation process was completed with a literature study and analysis of Gene Ontology (GO) annotations (The Gene Ontology Consortium, 2000) for genes ranked by our method and for genes selected by Dudoit *et al.* We will call these gene sets MC (Monte Carlo) and Dudoit, respectively.

The top-ranked genes by both MC and Dudoit analysis were separated in three groups: genes only found by MC, genes only found by Dudoit and genes found by both methods. After that, GO analysis was performed for each of the groups, using all the genes on the chips as background. We then focused on the biological process categories that were significantly different between MC and Dudoit's groups. There were numerous biological process categories overrepresented for each of the two groups. At first sight, the overrepresented categories in Dudoit group seem more relevant or cancer related, e.g. DNA recombination, spindle localization, cell cycle checkpoint, DNA metabolism or DNA repair. These categories include genes germane to different types of tumors, such as ATRX, IL7R, RAG2, ATR or RB1 (Dibirdik *et al.*, 1991; Gladdy *et al.*, 2003;

Melo *et al*., 1998; Nieborowska-Skorska *et al*., 2006; Xue *et al*., 2003). Furthermore, in some cases, a few of these genes have been preferentially associated to ALL but not AML, and vice versa.

On the other hand, GO analysis of MC shows over-represented categories with no immediate connection to cancer. However, there is a clear overrepresentation of a set of categories related to immune response and defense to biotic stimuli, e.g. chemokine biosynthesis and metabolism, defense response to bacteria, response to wounding, response to bacteria, monocyte activation, cytokine biosynthesis and metabolism, positive regulation of immune response, macrophage chemotaxis, complement activation, inflammatory response, etc. Most of these responses may be summarized as part of the innate immune response that, in fact, is also an overrepresented GO category itself. Macrophages/monocytes and Polymorphonuclear leucocytes (PMN)/granulocytes constitute the major cell types of the innate immune system, while lymphocytes (B and T cells) are the cellular components of the adaptive immune system. This is important to remember here, because the two types of acute leukemia being studied, AML and ALL, differ in the cell types that are affected. In Acute Myelogenous (granulocytic) Leukemia (AML), the leukemic cells are primarily of myeloid origin (granulocytes or monocytes), while in Acute Lymphocytic (lymphoblastic) Leukemia (ALL) cancerous cells arise from lymphocytes (B and T cells). Therefore, the primary cell origin for each type of leukemia is clearly distinct in their lineage. As mentioned before, MC analysis detected a clear overrepresentation of genes and categories involved in the innate response against bacteria and other biotic stimuli. This clearly indicates that there are subsets of genes characteristic of cells of myeloid origin that are good classifiers between AML and ALL.

There are several subtypes of AML, depending on the origin of the abnormal cells and their grade of differentiation. The most abundant type of granulocytes, the neutrophils and their precursors, are involved in most types of AML, although frequently there are also subtypes involving monocytes. Having these basic ideas and notions in mind, it is worth analyzing in more detail some of the genes and categories found over-represented in the MC analysis.

First, there were several genes differentially expressed and involved in proteolysis. Interestingly, several of these genes were found to be specifically neutrophil proteases (AZU1, CTSG, ELA2) (Wiedow and Meyer-Hoffert, 2005; Wong *et al*., 1999) that accumulate in azurophilic granules of neutrophils, playing a basic role in antimicrobial defense. Furthermore, AZU1 and ELA2 form a gene cluster with PRTN3 and DF, also selected by the MC analysis, which are all proteases (Wong *et al*., 1999). The chromatin in this cluster suffers reorganization during myeloid differentiation resulting in myeloid-specific transcription of the cluster. Moreover, these genes have previously been associated with myeloid leukemia and myeloid differentiation (Dunne *et al*., 2006; El-Ouriaghli *et al*., 2003; Lane and Ley, 2003). Second, there were genes (PFC and PTX3) involved in the response to external stimuli that seem to be preferentially expressed in monocytes (Doni *et al*., 2003; Polentarutti *et al*., 1998; Schwaeble *et al*., 1994). Thirdly, there were several genes occurring in the overrepresented categories related to the innate immune response that have previously been associated with myeloid differentiation, myeloid leukemia etiology and/or serve as markers of myeloid lineage. We could mentioned CD36 (Bordessoule *et al*., 1993; Perea *et al*., 2005) and CCL23 (Steinbach *et al*., 2006) that are over-expressed and serve as markers of myeloid leukemia, while other genes such as, for instance, the neutrophil chemokine CXCL2 (Belo *et al*., 2005), ANPEP (Alfalah *et al*., 2006) or IL18 (Robertson *et al*., 2006) are usually expressed by myeloid cells. Finally, most of the aforementioned genes were typically found to be highly expressed by myeloid-related cells (e.g. BM-CD33+ myeloid, PB-BDCA4+ dendritic cells, PB-CD14+ monocytes, PB-CD56+ NK cells and HL60) and very poorly expressed by cells of lymphocytic origin as found in the human UCSC genome browser microarray expression data.

In conclusion, the method of Dudoit *et al*. which selects genes that are basically displaying very different levels of expression between AML and ALL results in a group of classifiers that includes typical cancer genes, which previously have been involved in various types of cancer and that play basic roles in DNA repair and cell cycle. On the other hand, the MC analysis seems to select genes that relate more to the cellular origin of the primary cells that initiate the leukemia. It is particularly striking to observe the overrepresentation of genes preferentially expressed in neutrophils and monocytes that are clear indicators of the myeloid origin of AML leukemic cells. For details concerning the biological validation please see http://www.lcb.uu.se/papers/draminski/MCFS/.

4 CONCLUDING REMARKS

The algorithm proposed herein provides an effective and reliable method for ranking features according to their importance in supervised classification. Clearly, this algorithm is of general applicability, not limited to analyzing microarray gene expression data.

Reliability of the obtained ranking of features is a consequence of relying on the Monte Carlo approach, with sufficiently numerous and extensive resampling, as well as on using a sufficiently flexible classifier, namely a tree classifier. It also follows from the way in which we choose the number s of subsets of features while other parameters remain fixed. Here, we mean the requirement that the distance [Equation (3)] between successive rankings stabilizes at some acceptably low level. Seemingly, this last requirement has also proven to prevent masking of some features from becoming a problem. Thanks to the validation and confirmatory steps, statistical significance of the results is appraised as well.

It should be emphasized that the algorithm has been designed specifically to rank features with respect to their classification ability, not only to find those features that are important for classification.

Biological validation of the MC genes from literature and GO annotations further suggests that our MC ranking is not only preferred to other methods when the goal of feature selection is to rank the features according to their classification ability, but also, at least in some cases, that it selects features that are germane to the origins of the undergoing processes, in our case the genesis of leukemia. Indeed, at least in the case of

leukemia, we can say that our method seems to discover causes rather than effects.

We will continue validating this hypothesis extensively in our LCB-Data Warehouse (Ameur *et al.*, 2006). However, we wish to make our findings available already now so that it may be validated by a broader community.

Note added in proof: since the submission of the first version of the article, validity of all claims made has been confirmed on several data sets of biological or commercial origin, including transactional data from a major multinational FMCG company and geological data from oil wells operated by a major American oil company.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for providing valuable comments. M.D. and J.K. were partially supported by the Swedish project, Human Research Potential and the Socio-economic Knowledge Base: Access to Research Infrastructures, HPRI-CT-2001-00153. A.R.-I. and C.W. were partially supported by the Swedish Research Council, J.K. and S.E. were partially supported by The Wallenberg Foundation and The Swedish Foundation for Strategic Research.

Conflict of Interest: none declared.

REFERENCES

Alfalah,M. *et al.* (2006) A mutation in aminopeptidase N (CD13) isolated from a patient suffering from leukemia leads to an arrest in the endoplasmic reticulum. *J. Biol. Chem.*, **281**, 11894–11900.

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by expression profiling. *Nature*, **403**, 503–511.

Ameur,A. *et al.* (2006) The LCB Data Warehouse. *Bioinformatics*, **22**, 1024–1026.

Belo,A.V. *et al.* (2005) Murine chemokine CXCL2/KC is a surrogate marker for angiogenic activity in the inflammatory granulation tissue. *Microcirculation*, **12**, 597–606.

Bordessoule,D. *et al.* (1993) Immunohistological patterns of myeloid antigens: tissue distribution of CD13, CD14, CD16, CD31, CD36, CD65, CD66 and CD67. *Br. J. Haematol.*, **83**, 370–383.

Breiman,L. *et al.* (1984) *Classification and Regression Trees*. Wadsworth International Group, Monterey, CA.

Dibirdik,I. *et al.* (1991) Engagement of interleukin-7 receptor stimulates tyrosine phosphorylation, phosphoinositide turnover, and clonal proliferation of human T-lineage acute lymphoblastic leukemia cells. *Blood*, **78**, 564–570.

Doni,A. *et al.* (2003) Production of the soluble pattern recognition receptor PTX3 by myeloid, but not plasmacytoid, dendritic cells. *Eur. J. Immunol.*, **33**, 2886–2893.

Dudoit,S. and Fridlyand,J. (2003) Classification in microarray experiments. In Speed,T. (ed.), *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC Press, Boca Raton, pp. 93-158.

Dudoit,S. *et al.* (2002) Multiple hypothesis testing in microarray experiments. *Technical report 110*. In Division of Biostatistics. University of California, Berkeley.

Dudoit,S. *et al.* (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, **18**, 71–103.

Dunne,J. *et al.* (2006) siRNA-mediated AML1/MTG8 depletion affects differentiation and proliferation-associated gene expression in t(8;21)-positive cell lines and primary AML blasts. *Oncogene*.

El-Ouriaghli,F. *et al.* (2003) Clonal dominance of chronic myelogenous leukemia is associated with diminished sensitivity to the antiproliferative effects of neutrophil elastase. *Blood*, **102**, 3786–3792.

Gladdy,R.A. *et al.* (2003) The RAG-1/2 endonuclease causes genomic instability and controls CNS complications of lymphoblastic leukemia in p53/Prkdc-deficient mice. *Cancer Cell*, **3**, 37–50.

Golub,T.R. *et al.* (1999) Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Lane,A.A. and Ley,T.J. (2003) Neutrophil elastase cleaves PML-RARalpha and is important for the development of acute promyelocytic leukemia in mice. *Cell*, **115**, 305–318.

Melo,M.B. *et al.* (1998) Molecular analysis of the retinoblastoma (RB1) gene in acute myeloid leukemia patients. *Leuk. Res.*, **22**, 787–792.

Nieborowska-Skorska,M. *et al.* (2006) ATR-Chk1 axis protects BCR/ABL leukemia cells from the lethal effect of DNA double-strand breaks. *Cell Cycle*, **5**, 994–1000.

Perea,G. *et al.* (2005) Adverse prognostic impact of CD36 and CD2 expression in adult de novo acute myeloid leukemia patients. *Leuk. Res.*, **29**, 1109–1116.

Polentarutti,N. *et al.* (1998) Interferon-gamma inhibits expression of the long pentraxin PTX3 in human monocytes. *Eur. J. Immunol.*, **28**, 496–501.

Robertson,S.E. *et al.* (2006) Expression and alternative processing of IL-18 in human neutrophils. *Eur. J. Immunol.*, **36**, 722–731.

Schwaeble,W. *et al.* (1994) Expression of properdin in human monocytes. *Eur. J. Biochem.*, **219**, 759–764.

Simon,R. *et al.* (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst*, **95**, 14–18.

Smyth,G.K. *et al.* (2003) Statistical issues in cDNA microarray data analysis. In Brownstein,M.J. and Khodursky,A.B. (eds.), *Functional Genomics: Methods and Protocols. Methods in Molecular Bilogy*. Vol. 224, Humana Press, Totowa, NJ, pp. 111–136.

Speed,T. (ed.), (2003) *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC Press, Boca Raton.

Steinbach,D. *et al.* (2006) Identification of a set of seven genes for the monitoring of minimal residual disease in pediatric acute myeloid leukemia. *Clin. Cancer Res.*, **12**, 2434–2441.

Su,Y. *et al.* (2003) RankGene: identification of diagnostic genes based on expresion data. *Bioinformatics*, **19**, 1578–1579.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Tibshirani,R. *et al.* (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.*, **18**, 104–117.

Wiedow,O. and Meyer-Hoffert,U. (2005) Neutrophil serine proteases: potential key regulators of cell signalling during inflammation. *J. Intern. Med.*, **257**, 319–328.

Wittenn,I.H. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd edn. Morgan Kaufmann, San Francisco.

Wong,E.T. *et al.* (1999) Changes in chromatin organization at the neutrophil elastase locus associated with myeloid cell differentiation. *Blood*, **94**, 3730–3736.

Xue,Y. *et al.* (2003) The ATRX syndrome protein forms a chromatin-remodeling complex with Daxx and localizes in promyelocytic leukemia nuclear bodies. *Proc. Natl Acad. Sci. USA*, **100**, 10635–10640.