BIOINFORMATICS

Vol. 19 no. 14 2003, pages 1800–1807
DOI: 10.1093/bioinformatics/btg238

New algorithms for multi-class cancer diagnosis using tumor gene expression signatures

A. M. Bagirov*, B. Ferguson, S. Ivkovic, G. Saunders and J. Yearwood

Centre for Informatics and Applied Optimization, University of Ballarat, Ballarat 3353, Australia

Received on September 2, 2002; revised on December 19, 2002; accepted on March 4, 2003

ABSTRACT

Motivation: The increasing use of DNA microarray-based tumor gene expression profiles for cancer diagnosis requires mathematical methods with high accuracy for solving clustering, feature selection and classification problems of gene expression data.

Results: New algorithms are developed for solving clustering, feature selection and classification problems of gene expression data. The clustering algorithm is based on optimization techniques and allows the calculation of clusters step-by-step. This approach allows us to find as many clusters as a data set contains with respect to some tolerance. Feature selection is crucial for a gene expression database. Our feature selection algorithm is based on calculating overlaps of different genes. The database used, contains over 16 000 genes and this number is considerably reduced by feature selection. We propose a classification algorithm where each tissue sample is considered as the center of a cluster which is a ball. The results of numerical experiments confirm that the classification algorithm in combination with the feature selection algorithm perform slightly better than the published results for multi-class classifiers based on support vector machines for this data set.

Availability: Available on request from the authors.

Contact: a.bagirov@ballarat.edu.au

1 INTRODUCTION

The study of specific methods for cancer diagnosis is very important. Different methods can be used for this purpose. One of them is molecular diagnostics which offers the promise of precise, objective, and systematic cancer classification, but these tests are not widely applied because characteristic molecular markers for most solid tumors have yet to be identified [see Connolly *et al*. (1997)]. The second method is the use of DNA microarray-based tumor gene expression profiles. Recently, they have been used for cancer diagnosis. In Alizadeh *et al*. (2000), Bittner *et al*. (2000), Dhanasekaran *et al*. (2001), Golub *et al*. (1999), Hedenfalk (2001), Perou *et al*. (2000) studies have been limited to a few cancer types and have spanned multiple technology platforms complicating comparison among different data sets. In Ramaswamy *et al*. (2001) a support vector machines (SVM) algorithm has been applied to solving the classification of tumors based on gene expression data gathered from microarray analysis.

This paper is based on the same data that are used in Ramaswamy *et al*. (2001). The distinctive features of this database are the very large number of attributes and the extremely small number of records. For example, many classes of tumors contain only eight samples in the training set. Therefore effective feature selection algorithms are crucial for this type of database. The small number of samples requires the use of very specific methods for classification.

Investigating the formation of clusters allows us to understand the structure of the database under consideration. It also raises questions about cancer classes as the results show clusters that correspond well with tumor classes and others which are rather mixed.

In this paper new clustering, feature selection and classification algorithms for gene expression databases are presented. We propose a clustering algorithm which calculates clusters step-by-step and calculates as many clusters as the database contains with respect to some tolerance. We develop a new feature selection algorithm which is particularly suitable for this multi-class problem with gene expression data. This algorithm essentially uses the overlaps for gene expression between different classes. We also propose an algorithm for solving the classification problem. In this algorithm we cover each class with balls. We take each tissue sample with gene expression profile from the training set as the centroid of a ball and calculate its radius using the distance between the sample and other classes. A new sample is assigned to a class with the least distance between the sample and the cover of this class. The results of numerical experiments confirm that this algorithm in combination with our feature selection algorithm improves the results achieved in Ramaswamy *et al*. (2001) using SVM and other feature selection techniques.

* To whom correspondence should be addressed.

2 ALGORITHMS

2.1 The database

This paper is based on data which was gathered from snap frozen human tumor and normal tissue specimens, spanning 14 different tumor classes. The 14 tumor classes were breast adenocarcinoma, prostate, lung adenocarcinoma, leukemia, colorectal adenocarcinoma, lymphoma, bladder, melanoma, uterine adenocarcinoma, renal cell carcinoma, pancreatic adenocarcinoma, ovarian adenocarcinoma, pleural mesothelioma and central nervous system. They were obtained from the National Cancer Institute/Cooperative Human Tissue Network, the Massachusets General Hospital Tumor Bank, the Dana-Farber Cancer Institute, Brigham and Women's Hospital, the Children's Hospital (all in Boston, USA), and the Memorial Sloan Kettering Cancer Center (New York, USA). All tumors were biopsy specimens from primary sites (except where noted) obtained before any treatment and were enriched in malignant cells ($>50\%$) but otherwise unselected. Normal tissue RNA was from snap-frozen autopsy specimens collected through the International Tissue Collection Network.

Hybridization targets were prepared with RNA from whole tumors by using methods described in Golub *et al*. (1999). Targets were hybridized sequentially to oligonucleotide microarrays containing a total of 16 063 probe sets representing 14 030 GenBank and 475 The Institute for Genomic Research accession nos., and arrays were scanned by using Standard Affymetrix protocols and scanners. Then each probe set was considered as a separate gene. Expression values for each gene were calculated by using Affymetrix GENECHIP analysis software.

Of 314 tumor and 98 normal tissue samples processed, 218 tumor and 90 normal tissue samples passed quality control criteria and were used for the data analysis. The resulting data set contains almost five million gene expression values. This database is publicly available at: www-genome.wi.mit.edu/MPR/GCM.html [see, also Ramaswamy *et al*. (2001)].

2.2 Tissue clustering

In this subsection we describe an algorithm for finding clusters in a data set. Different algorithms for clustering can be found in Hawkins *et al*. (1982), Spath (1980) and Jain *et al*. (1999). The formulation of k-means clustering as a mathematical programming problem has been developed in Bradley *et al*. (1999). We present a formulation of the clustering problem in terms of nonsmooth and nonconvex optimization. This formulation can also be found in Bagirov *et al*. (2001, 2002a,b).

Consider a set A that consists of n p-dimensional tissue samples with gene expression profile $a^i = (a^i_1, \dots, a^i_p)$, $i = 1, \dots, n$. The aim of clustering is to represent this set as the union of q clusters. Since each cluster can be described by its centroid, we would like to find q cluster centroids.

Consider now an arbitrary set X, consisting of q cluster centroids $x^1, \dots, x^q$. The distance $d(a^i, X)$ from a tissue sample $a^i \in A$ to this set is defined by

$$d(a^i, X) = \min_{s=1,\dots,q} \|x^s - a^i\|.$$

Here $\|\cdot\|$ is Euclidean norm.

The deviation $d(A, X)$ from the set A to the set X can be calculated using the formula

$$d(A, X) = \sum_{i=1}^{n} d(a^i, X) = \sum_{i=1}^{n} \min_{s=1,\dots,q} \|x^s - a^i\|.$$

The deviation is the sum of distances of each tissue to the adjacent cluster centroid. Thus, as far as the optimization approach is concerned, the cluster analysis problem can be reduced to the following problem of mathematical programming which finds the cluster centroids minimizing the sum of the deviations of all tissue samples:

$$\begin{aligned} &\text{minimize} \quad f(x^1, \dots, x^q) \\ &\text{subject to} \quad (x^1, \dots, x^q) \in R^{p \times q}, \end{aligned} \tag{1}$$

where

$$f(x^1, \dots, x^q) = \sum_{i=1}^{n} \min_{s=1,\dots,q} \|x^s - a^i\|. \tag{2}$$

If $q > 1$, the objective function (2) in the problem (1) is nonconvex and nonsmooth. The number of variables in this problem is $q \times p$. If the number q of clusters and the number p of attributes are large, the decision maker is facing a large-scale global optimization problem. Moreover, the form of the objective function in this problem is complex enough not to become amenable to the direct application of general purpose global optimization methods. Therefore, in order to ensure the practicality of the optimization approach to clustering, the proper identification and use of local optimization methods with an appropriate choice of starting point is very important. Clearly, such an approach does not guarantee a globally optimal solution to problem (1). On the other hand, this approach will find a local minimum of the objective function that, in turn, provides a good enough clustering description of the data set under consideration.

Note also that a meaningful choice of the number of clusters is very important for cluster analysis. It is difficult to define *a priori* how many clusters represent the set A under consideration. The following strategy can be used here: starting from a small enough number of clusters q, the decision maker has to gradually increase the number of clusters for the analysis until certain termination criteria motivated by the underlying decision making situation are satisfied.

From an optimization perspective this means that if the solution of the corresponding optimization problem (1) is not satisfactory, the decision maker needs to consider the problem (1) with $q + 1$ clusters and so on. This implies that one needs to repeatedly solve global optimization problems of type (1) with different values of q—a task even more challenging than solving a single global optimization problem. In order to avoid this difficulty, a step-by-step calculation of clusters is implemented in the algorithm discussed below.

2.2.1 An optimization clustering algorithm

ALGORITHM 1. Clustering.

Step 1. (Initialization). Select a tolerance $\varepsilon > 0$. Select a starting cluster centroid $x^0 = (x_1^0, \dots, x_p^0) \in R^p$ and solve the minimization problem (1). Let $x^{1*} \in R^p$ be a solution to this problem and f^{1*} be the corresponding objective function value. Set $k = 1$.

Step 2. (Identification of the next cluster). Select a starting cluster centroid $x^0 \in R^p$, and solve the following minimization problem:

$$\begin{aligned} &\text{minimize} \quad \bar{f}^k(x) \\ &\text{subject to} \quad x \in R^p \end{aligned} \tag{3}$$

where

$$\bar{f}^k(x) = \sum_{i=1}^{n} \min\{\|x^{1*} - a^i\|, \dots, \|x^{k*} - a^i\|, \|x - a^i\|\}.$$

Step 3. (Refitting of all clusters). Let $\bar{x}^{k+1,*}$ be a solution to the problem (3). Take $x^{k+1,0} = (x^{1*}, \dots, x^{k*}, \bar{x}^{k+1,*})$ as a new starting cluster centroids and solve the following minimization problem:

$$\begin{aligned} &\text{minimize} \quad f^{k+1}(x) \\ &\text{subject to} \quad x \in R^{(k+1)\times p} \end{aligned} \tag{4}$$

where

$$f^{k+1}(x) = \sum_{i=1}^{n} \min_{j=1,\dots,k+1} \|x^j - a^i\|.$$

Step 4. (Stopping criterion). Let $x^{k+1,*}$ be a solution to the problem (4) and $f^{k+1,*}$ be the corresponding value of the objective function. If

$$\frac{f^{k*} - f^{k+1,*}}{f^{1*}} < \varepsilon$$

then stop, otherwise set $k = k + 1$ and go to Step 2.

In Step 1 the centroid of the set A is calculated. In this step the problem of convex programming is solved. In Step 2 we calculate the centroid of the next $(k + 1)th$ cluster. In these two steps the number of variables in the corresponding optimization problems is p which is substantially less than that in the original problem (1). In Step 3 we refine all cluster centroids. Such an approach allows one to significantly reduce the computational time for solving problem (4). It can be shown that $f^{k*} \geq f^{k+1,*} \geq 0$ for all $k \geq 0$. Thus as a result we will have a decreasing sequence $\{f^{k*}\}$ and $f^{k*} \geq 0$ for all $k \geq 0$. The latter implies that after $\bar{k}$ iterations the stopping criterion in Step 4 will be satisfied.

Problems (3) and (4) in Steps 2 and 3, respectively, are problems of global optimization. Since the number of variables in these problems are large the global optimization techniques fail to solve them. Therefore it is very important to use methods which can find local minima providing good cluster description of the data set under consideration. The *discrete gradient method* described in Bagirov (1999) is one such method. The discrete gradient method is a method of nonsmooth optimization where subgradients are replaced by their approximations—discrete gradients. The discrete gradient is a finite-difference estimate of a subgradient and is calculated with respect to a given direction using some step in this direction. A terminating algorithm for the calculation of a descent direction of an objective function is proposed in Bagirov (1999). Since for stationary points which are not local minima there always exists a descent direction, this algorithm finds such a direction and escapes from these points. On the other hand if the size of the valley where the local minimum is located is small enough then large enough values of the step for the calculation of the discrete gradient may allow escape from such a local minimum. Results from Bagirov and Rubinov (2003) confirm this. But it should be noted that this is not always true. Cluster centroids in the proposed clustering algorithm are calculated step by step and previous cluster centroids are used to find the centroid of the next cluster. Such a strategy together with the properties of the discrete gradient method allow the determination of a local minimum which provides a good description of a data set.

The choice of parameter $\varepsilon > 0$ is very important in Algorithm 1. Large values of ε can lead to big clusters which are the union of other clusters whereas small values of ε can lead to the appearance of small and artificial clusters. The choice of ε is discussed in Bagirov and Yearwood (2003, submitted for publication). (Available on: http://www.ballarat.edu.au/itms/research_papers/papers2003.shtml.) It was noted that the best values for ε are $\varepsilon \in [0.01, 0.1]$. The values $\varepsilon < 0.01$ lead to the appearance of artificial clusters.

It should be noted that the optimization is impractical with the full set of genes, so the proposed clustering algorithm requires a preliminary dimension reduction and is not directly applicable to gene expression data. Our numerical experience shows that the maximum number of genes for which the proposed clustering algorithm can be effectively run is of the order of 800 genes. This algorithm can run more effectively when the number of genes is restricted to 350, otherwise the calculation of clusters requires too much CPU time.

2.3 Feature selection algorithm

In this subsection we describe a feature selection algorithm. The gene expression database under consideration contains 16 063 genes and the results of numerical experiments show that many of them are not informative for solving the multi-classification problem.

The preliminary analysis of microarray data show that for each cancer type there exists a subset of genes which is responsible for this type or can be used for its better description. The intervals where these genes change for one particular cancer type has no or almost has no intersections with the intervals for this gene in many other cancer types. Therefore if we compare two cancer types we can find quite large subset of genes for which these intervals have empty or almost empty intersection. Then these genes can be used for the discrimination of these two cancer types. Since number of cancer types is large enough it is very difficult to find a subset of genes which are very good for the discrimination of all cancer types. Therefore it is very important to find genes which are very good for the discrimination of as many cancer types as possible. The feature selection algorithm described in this section tries to determine such genes.

Our feature selection algorithm is based on the overlaps of gene expression values between different classes.

Suppose there are m classes and n_i tumors in the ith class. Let p be the number of features in the data set. We introduce the following numbers:

$$a_{ij}^{\min} = \min_{k=1,\dots,n_i} d_{kj}^i, \quad a_{ij}^{\max} = \max_{k=1,\dots,n_i} d_{kj}^i,$$
$$j = 1, \dots, p, \quad i = 1, \dots, m.$$

where d_{kj}^i is the jth gene expression value for kth tumor in the ith class. Here $a_{ij}^{\min}(a_{ij}^{\max})$ is the minimum(maximum) value for jth gene in ith class. Thus the jth feature in the ith class can be identified by a segment $[a_{ij}^{\min}, a_{ij}^{\max}]$. For a given gene $j = 1, \dots, p$ and two different classes i_1 and i_2 we define the following quantity:

$$O_{i_1,i_2}^j = 1$$

if $[a_{i_2j}^{\min}, a_{i_2j}^{\max}] \subset [a_{i_1j}^{\min}, a_{i_1j}^{\max}]$ or $[a_{i_1j}^{\min}, a_{i_1j}^{\max}] \subset [a_{i_2j}^{\min}, a_{i_2j}^{\max}]$, otherwise

$$O_{i_1,i_2}^j = \frac{b_1}{b_2}$$

where

$$b_1 = \min(a_{i_1j}^{\max}, a_{i_2j}^{\max}) - \max(a_{i_1j}^{\min}, a_{i_2j}^{\min}),$$
$$b_2 = \max(a_{i_1j}^{\max}, a_{i_2j}^{\max}) - \min(a_{i_1j}^{\min}, a_{i_2j}^{\min}).$$

The quantity O_{i_1,i_2}^j is said to be the overlap of jth gene between classes i_1 and i_2. It is clear that $O_{i_1,i_2}^j = O_{i_2,i_1}^j$ and $O_{i_1i_2}^j \le 1$ for any $j = 1, \dots, p$ and $i_1, i_2 = 1, \dots, m$. It should be noted that the overlap $O_{i_1i_2}^j$ can be negative. Figure 1 illustrates the overlap for two genes between two classes. The overlap for gene 1 is $(a_{11}^{\max} - a_{21}^{\min})/(a_{21}^{\max} - a_{11}^{\min})$ and for gene 2 it is $(a_{22}^{\max} - a_{12}^{\min})/(a_{12}^{\max} - a_{22}^{\min})$. Gene 1 is more discriminative than gene 2.

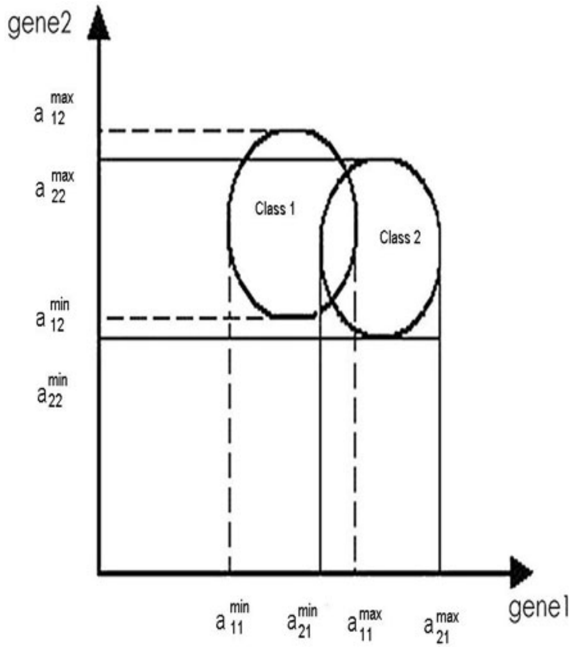

Fig. 1. Gene overlap between two classes.

ALGORITHM 2. The overlap algorithm for feature selection.

Step 1. (Initialization). Let $j = 0$, $\alpha < 1$ and $c_0 > 0$ be an integer number.

Step 2. Set $j = j + 1$. If $j > p$ then stop. Otherwise go to Step 3.

Step 3. Calculate overlaps O_{ik}^j for all $i, k = 1, \dots, m$.

Step 4. Set $c(j) = 0,\ i = 0$.

Step 4a. Set $i = i + 1$. If $i \ge m$ then go to Step 2, otherwise go to Step 4b.

Step 4b. Set $k = i + 1$. If $k \ge m$ go to Step 4a, otherwise go to Step 5.

Step 5. If $O_{ik}^j \le \alpha$ then set $c(j) = c(j) + 1$. If $c(j) \ge c_0$ go to Step 2, otherwise go to Step 4b.

Step 6. Remove all j with $c(j) < c_0$ and create a new set of features with $c(j) = c_0$. This new set of features is a set of informative features.

The algorithm contains parameters which deserve some explanation as the choice of parameters $\alpha < 1$ and $c_0 > 0$ is very important. The parameter α indicates the degree of overlap. If $\alpha \le 0$ for overlap $O_{i_1i_2}^j$ the intersection of segments $[a_{i_1j}^{\min}, a_{i_1j}^{\max}]$ and $[a_{i_2j}^{\min}, a_{i_2j}^{\max}]$ is empty for the gene j which in its turn means that the jth gene is informative for the discrimination of the classes i_1 and i_2. Larger values of α

means that this intersection is big enough and the jth gene is not informative for the discrimination of these classes. It is quite possible that for $\alpha = 0$ or small values of α the set of informative features can be empty. Therefore we should consider different values for α to get the most informative genes and the appropriate value of α depends on the database under consideration. The variable $c(j)$ contains the number of pairs of classes for which overlaps for the jth gene are less than the number α, that is the jth gene is 'good' for their discrimination. The maximum number of these pairs is $l = m(m-1)/2$. So $c_0 \le l$. If the number of classes in the database is $m = 2$, then $l = 1$ and in this case $c_0 = l$. If m is large enough then c_0 cannot be very close to l because it is very difficult to find a gene which will be useful for the discrimination of all pairs of classes.

Let

$$\alpha_0 = \min_{i_1,i_2,j} O^j_{i_1 i_2}.$$

For values of α close to α_0 the values of c_0 have to be small. If α is much greater than α_0 the value of c_0 can be taken large enough, but less than l.

2.4 Classification

k-NN algorithm is one of the fast and effective algorithms for data classification. In this subsection we propose an algorithm which can be considered as a modification of k-NN algorithm.

In describing the classification algorithm, we suppose that the database A contains m classes $A_i, i = 1, \dots, m,\ m \ge 2$. Let $\beta \in [0, 0.5]$ be a given number. First for all $x \in A_1$ we calculate the distance between this tissue sample and the union of all other classes $A_i,\ i = 2, \dots, m$:

$$\delta_1(x) = \min \left\{ \|x - y\| : y \in \bigcup_{i=2}^{m} A_i \right\}.$$

Then for $x \in A_1$ we define:

$$\sigma_1(x) = \beta\delta_1(x).$$

We consider a ball $S_{\sigma_1(x)}(x)$ with the center at the tissue sample $x \in A_1$ and radius $\sigma_1(x)$ and construct the following set:

$$\bar{A}_1 = \bigcup_{x \in A_1} S_{\sigma_1(x)}(x).$$

It is clear that

$$A_1 \subset \bar{A}_1,$$

and

$$\bar{A}_1 \bigcap A_i = \emptyset, \quad i = 2, \dots, m.$$

Assume that we have already constructed the sets $\bar{A}_j,\ j = 1, \dots, k,\ k < m$. The next set $\bar{A}_{k+1}$ is constructed as follows. For all $x \in A_{k+1}$ we calculate:

$$\delta_{k+1}(x) = \min \left\{ \|x - y\| : y \in \bigcup_{i=1, i \ne k+1}^{m} A_i \right\}.$$

We take $\sigma_{k+1}(x) = \beta\delta_{k+1}(x)$ and consider a ball $S_{\sigma_{k+1}(x)}(x)$ with the center at $x \in A_{k+1}$ and radius $\sigma_{k+1}(x)$. Then we construct the following set:

$$\bar{A}_{k+1} = \bigcup_{x \in A_{k+1}} S_{\sigma_{k+1}(x)}(x).$$

We calculate all other sets $\bar{A}_j, j = k+2, \dots, m$ in the same way. Note that by construction

$$\bar{A}_i \bigcap \bar{A}_k = \emptyset \quad \text{for all } i, k = 1, \dots, m,\ \ i \ne k.$$

For a new tissue sample $x \in R^p$ we calculate the distance between this tissue sample and all sets $\bar{A}_j,\ j = 1, \dots, m$ and identify this tissue sample with the set to which it is closest. We will call this algorithm *the covering classification algorithm* (*CCA*). Note that this algorithm is invariant against a re-ordering of the classes.

We can see that if $\beta = 0$, the covering classification algorithm coincides with the k-NN algorithm with $k = 1$. The covering classification algorithm allows us to take into account the structure of the database under consideration using the distance between classes. Since $\beta \in [0, 0.5]$ this algorithm can be considered as a modification of the k-NN algorithm. Results of experiments presented in Section 4 show that this algorithm performs better than k-NN.

3 IMPLEMENTATION

In this section we describe the conditions under which our clustering, feature selection and classification algorithms have been applied to a gene expression database. The first task in dealing with the data under consideration was to normalize the features. This was done by linear transformations so that the mean value of all features were 1.

3.1 Clustering

The objective functions of the clustering algorithm in both problems (3) and (4) are nonsmooth and we apply the *discrete gradient* method from Bagirov (1999) to solve them. As was mentioned above this method overcomes stationary points which are not local minima. On the other hand the calculation of the clusters step-by-step in Algorithm 1 allows the determination of local minima of problem (4) which gives, as a rule, a good cluster description of the data set under consideration.

In Step 4, (*stopping criterion*), we take $\varepsilon = 0.01$ in order to allow the algorithm to calculate as many clusters as possible, knowing that the database contains quite large number of classes.

The clustering algorithm was applied to the database with 144 tumors which span the 14 common human cancer types listed earlier. Since the optimization based clustering algorithm is impractical when all genes are present we used the following method to reduce the number of genes. The standard deviation of each gene was calculated and all genes with standard deviation less than some threshold were removed. Different values of this threshold led to different subsets of features. We have considered different subsets of features from 200 to 800 genes. We used the following strategy for the calculation of clusters. First we apply the clustering algorithm to entire data set using $\varepsilon = 0.01$ in Step 4. Then we apply this algorithm to a cluster if this cluster contains more than 50% of all tissue samples. Then we join all small clusters which contain 2–3 tissue samples to the closest clusters.

3.2 Classification

For the feature selection algorithm we took $c_0 \in [15, 45]$ and $\alpha \in [0, 0.25]$. For the database under consideration $l = 91$ and larger values for c_0 led to the empty sets of features. Reasonable values for c_0 depend on the value of α. For any α there exists $c_1 \in [0, l]$ such that for any $c_0 > c_1$ the corresponding subset will be empty. Therefore we cannot take c_0 very large. The feature selection was performed on the training set. The algorithm significantly reduces the number of features. Different subsets of features were considered which contain from 200 to 700 genes.

In our classification algorithm the only parameter is $\beta \in [0, 0.5]$. We find this parameter using the training set. A portion, say 2/3 of the training set taken randomly is considered as a new training set and the remaining part as a test set. Then we consider the following values for $\beta = 0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.5$ and choose a value which provides best accuracy for the test set. We repeat this three times and define a value of β as an average of best values.

4 RESULTS

In this section we present the results obtained from the clustering, feature selection and classification algorithms applied to the gene expression database.

4.1 Clustering

The results obtained by the clustering algorithm were as follows. We calculated three subsets of features. These feature subsets consisted of 286, 336 and 737 genes, respectively. These subsets of genes were obtained by using standard deviations. It should be noted that there was no big difference between the results for 336 and 737 genes. Further decreases in the number of genes leads to changes in the cluster structure of the database. So we can conclude that there exist 280–350 genes which determine the cluster structure of the database and all other genes play a less informative role. The clustering algorithm calculated 14 different clusters. Seven of them correspond directly to tumor classes as they contain only one type of tumor. Lymphoma, leukemia and CNS have their own clusters which contain almost all of these kinds of tumors. Prostate, melanoma, uterus and mesothelioma also have their own clusters, but these clusters do not contain all of these tumors. There are two other clusters where one of the tumor types dominate. These clusters contain mostly breast cancer and colorectal. The other five clusters are a mixture of different tumor types. The overall accuracy (in terms of correspondence with tumor classes) of the clustering algorithm for 336 and 737 genes was 63.2%, however for 286 genes it was 61.9%.

Table 1. Results for different number of genes: the multiclass problem

No. of genes	1-NN	CCA
20	48.2	48.2
69	48.2	48.2
158	64.8	68.5
255	74.1	75.9
324	74.1	79.6
403	70.4	72.2
613	68.5	72.2
16 063	48.2	59.3

We applied the k-means algorithm to find 14 clusters in this data set using the same subsets of 286, 336 and 737 genes. The results are as follows. Leukemia has its own cluster which contains most of this cancer type. Breast cancer, prostate, uterus, pancreas and CNS have clusters which contain only those cancer types, but not all of them. Other clusters are a mixture of different cancer types. The overall accuracy of k-means was 48.61, 45.8 and 44.44% for the subsets of 286, 336 and 737 genes, respectively. Further increasing the number of clusters did not lead to significant improvements. We can see that the new clustering algorithm considerably improves the results obtained by k-means algorithm.

4.2 Classification

Two kinds of classification problems have been considered in this gene expression database. First 198 cancer tumors (144 of them as a training set and the remaining 54 as a test set) were used in order to diagnose different types of cancer tumors. We applied the overlap feature selection algorithm to the training set to reduce the number of genes. Different values of α and c_0 lead to different subsets of genes. More stable results were obtained by using subsets which contain from 250 to 700 genes. Both the k-NN and the covering classification algorithms were applied to the data. Accuracy is defined as the percentage of well classified tissue samples on the test set. The accuracy for k-NN was between 68 and 74% whereas our covering algorithm achieved accuracy from 72 to 80%. It should be noted that we used k-NN with $k = 1$, so it is quite possible that k-NN might yield better results if k and the

Table 2. Confusion matrix for the test set

	BR (%)	PR (%)	LU (%)	CO (%)	LY (%)	BL (%)	ML (%)	UT (%)	LE (%)	RE (%)	PA (%)	OV (%)	ME (%)	CNS (%)	*n*
BR	25		25			50									4
PR	16.6	66.7						16.6							6
LU			100												4
CO				75				25							4
LY					100										6
BL						33.3		33.3			33.3				3
ML							100								2
UT								100							2
LE									100						6
RE	33.3							33.3		33.3					3
PA				33.3							66.7				3
OV												100			4
ME													100		3
CNS														100	4
n	3	4	5	4	6	3	2	6	6	1	3	4	3	4	54

number of genes were tuned. The best result was obtained with the subset containing 324 genes. This subset was obtained by using the overlap feature selection algorithm with $\alpha = 0.22$ and $c_0 = 39$. The value for β was 0.15. It should be noted that k-NN with $k = 1$ and the covering classification algorithm using all 16 063 genes achieved 48 and 60% accuracy, respectively. Results for $\alpha = 0.22$ with different values of c_0 and consequently for different numbers of genes are presented in Table 1. These results confirm that feature selection for the database is crucial and there exist 200–400 genes which provide better diagnostic accuracy for cancer tumors. A confusion matrix for the test set is presented in Table 2. From this table we can see that lung, lymphoma, melanoma, uterus, leukemia, ovary, mesothelioma and CNS have been classified 100%, colorectal 75%, prostate and pancreas 67%, breast cancer 25%, bladder and renal 33% in testing phase. We can see that cancer types for which a clear cluster structure was discovered by the clustering algorithm have been classified more accurately.

Table 3. Results for different number of genes: the binary problem

No. of genes	1-NN	CCA
6	71.8	72.2
15	73.9	73.9
47	73.2	78.2
58	86.8	89.3
67	86.1	87.9
76	88.6	92.2
80	90.4	92.5
83	90.4	92.2
89	90.0	92.2
103	88.6	91.8
141	91.4	92.5
205	86.8	90.4
251	87.1	92.2
360	86.8	91.1
782	88.2	92.5
1772	86.8	89.6
16 063	76.4	76.7

4.3 Discriminating malignant and normal tissue

The second problem was the discrimination of malignant and normal tissues. We used 190 tumor tissues and 90 normal tissues—8 tumors which were re-occurrence cases were excluded. We used the ‘leave one out’ method to define the accuracy of different algorithms. Results of numerical experiments are summarized in Table 3. Since we have only one pair of classes $c_0 = 1$ in the feature selection algorithm. In each cycle, the feature selection algorithm was applied for the same value of α. For example, for the subset of 80 genes $\alpha = 0.1$. Then the classification algorithm was applied using the subsets of genes calculated from the overlap feature selection algorithm. The number of these genes can be slightly different in different cycles. In Table 3 we present their average. Results presented in Table 3 show that there exists subsets of 80–90 genes which provide better classification accuracy for the binary problem. Moreover the increase of the number of genes do not improve classification accuracy. For the subset of 80 genes the k-NN algorithm achieved 90.4% and the covering classification algorithm proposed in this paper achieved 92.5% accuracy for the test set. These algorithms achieved 76.4 and 76.7% accuracy using all genes. We again see that feature selection algorithm significantly reduced the number of genes and improved the results of the classification algorithm.

5 CONCLUSIONS

We have considered clustering, feature selection and classification problems in a gene expression database. Special algorithms have been proposed which seem to be more suitable for the database under consideration. The clustering algorithm which is based on methods of nonsmooth optimization, calculates clusters step-by-step and allows the calculation of as many clusters as the database contains up to a certain tolerance. The results of numerical experiments using 144 cancer tumors show that some cancer types have a clear structure in a gene expression space whereas other types of cancer have no such structure. Furthermore this algorithm considerably improved results obtained by using the k-means algorithm.

We have developed a feature selection algorithm which is based on the overlaps for a given gene between different classes. There is some similarity between this and an algorithm developed in Park *et al*. (2001). However this algorithm uses reordering of genes instead of selection based on overlaps in gene expression values between classes. The comparative analysis of these two algorithms are very interesting and will be subject of our further research.

We have considered two kind of classification problems in this database: discrimination of different tumor classes and discrimination of malignant and normal tissues. In both cases first we applied feature selection algorithms. Results of our calculations show that in the first case 2–3% percent of all genes are enough for the best classification of tumor classes. In the second case less than 1% of all genes provides the best discrimination of malignant and normal tissues.

In particular the classification results for the gene-based classification algorithm presented in this paper in combination with the overlap feature selection technique using far fewer genes achieve classification accuracy of 80% which is slightly better than the classification accuracy of 78% achieved by the SVM approach in Ramaswamy *et al*. (2001). The result for the binary problem was 92.5% which is again slightly better than the classification accuracy of 92% achieved by the SVM [see Ramaswamy *et al*. (2001)]. However, the increase in performance might be due to the combination of the classification and feature selection algorithms.

ACKNOWLEDGEMENTS

The authors would like to thank the three anonymous referees whose very detailed comments have considerably improved this paper.

REFERENCES

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

Bagirov,A.M. (1999) Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices. In Eberhard,A. *et al*. (eds.) *Progress in Optimization: Contribution from Australasia*. Kluwer Academic Publishers, Dordrecht, pp. 147–175.

Bagirov,A.M. and Rubinov,A.M. (2003) Cutting angle method and a local search. *J. Global Optimization*, to appear.

Bagirov,A.M., Rubinov,A.M. and Yearwood,J. (2001) Using global optimization to improve classification for medical diagnosis and prognosis. *Topics in Health Inform. Management*, **22**, 65–74.

Bagirov,A.M., Rubinov,A.M. and Yearwood,J. (2002a) A heuristic algorithm for feature selection based on optimization techniques. In Sarker,R., Abbas,H. and Newton,C.S. (eds.), *Heuristic and Optimization for Knowledge Discovery*, Idea Publishing Group, Hershey, pp. 13–26.

Bagirov,A.M., Rubinov,A.M. and Yearwood,J. (2002b) A global optimization approach to classification. *Optimization Eng.*, **3**, 129–155.

Bittner,M. *et al*. (2000) Molecular classification of cutaneous malignant melonoma by gene expression profiling. *Nature*, **406**, 536–540.

Bradley,P.S. *et al*. (1999) Mathematical programming for data mining: formulations and challenges. *INFORMS J. Comput.*, **11**, 217–238.

Connolly,J.L. *et al*. (1997) Principles of Cancer Pathology. In Holland,J.F. *et al*. (eds.), *Cancer Medicine*, Williams and Wilkins, Baltimore, pp. 533–555.

Dhanasekaran,S.M. *et al*. (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**, 822–826.

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

Hair,J.F. *et al*. (1998) *Multivariate Data Analysis*. Prentice-Hall, Englewood Cliffs, NJ.

Hedenfalk,I. *et al*. (2001) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, **344**, 539–548.

Hawkins,D.M. *et al*. (1982) Cluster analysis. In Hawkins,D.M. (ed.) *Topics in Applied Multivariate Analysis*. Cambridge University press, Cambridge.

Jain,A.K. *et al*. (1999) Data clustering: a review. *ACM Comput. Surveys*, **31**(3), 264–323.

Park,P. *et al*. (2001) A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pasific Symposium on Biocomputing*, vol 6, pp. 52–63.

Perou,C.M. *et al*. (2000) Molecular portraits of human breast tumors. *Nature*, **406**, 747–752.

Ramaswamy,S. *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci.*, **98**, 15149–15154.

Spath,H. (1980) *Cluster Analysis Algorithms*. Ellis Horwood Limited, Chichester.