A global optimisation approach to classification in medical diagnosis and prognosis*

Adil Bagirov, Alex Rubinov and John Yearwood
School of Information Technology and Mathematical Sciences
University of Ballarat
Ballarat, Australia
abaghirov@students.ballarat.edu.au, a.rubinov, j.yearwood@ballarat.edu.au
Andrew Stranieri
Department of Computer Science
La Trobe University
Bundoora, Australia
stranier@latcs1.cs.latrobe.edu.au

Abstract

In this paper global optimisation-based techniques are studied in order to increase the accuracy of medical diagnosis and prognosis with FNA image data from the Wisconsin Diagnostic and Prognostic Breast Cancer databases. First we discuss the problem of determining the most informative features for the classification of cancerous cases in the databases under consideration. Then we apply a technique based on convex and global optimisation to breast cancer diagnosis. It allows the classification of benign cases and malignant ones and the subsequent diagnosis of patients with very high accuracy. The third application of this technique is a method that calculates centres of clusters to predict when breast cancer is likely to recur in patients for which cancer has been removed. The technique achieves higher accuracy with these databases than reported elsewhere in the literature.

1. Introduction

In the last decade many approaches to classification have been used for health care applications. In particular, in breast cancer diagnosis and prognosis (for example, [15, 29, 28, 27]) discrimination approaches have been used. A woman’s chances for long term survival are improved by early detection of the cancer, and early detection is enhanced by accurate diagnosis techniques. Most breast cancers are detected by the patient or by screening as a lump in the breast. In order to diagnose whether the lump is benign or malignant, the physician may use mammography, FNA (Fine Needle Aspirate) directly from the breast lump with visual interpretation or surgical biopsy. The reported ability to correctly diagnose cancer when the disease is present is between 68% - 79% for mammography, 65% - 98% for FNA and close to 100% for surgical biopsy.

This paper describes the application of a new classification algorithm for diagnosing cancer based on measurements of the nuclei of the constituent cells [24] from the FNA fluid. A typical image may contain 10 to 40 nuclei and the image analysis process would take only seconds. Ten real-valued features are computed for each cell nucleus: radius (mean of distances from center to points on the perimeter), texture (standard deviation of grey-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness, concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension (of the boundary). The mean value, standard error and extreme value (i.e., largest or worst value: biggest size, most irregular shape) of each of these cellular features are computed for each image, resulting in a total of 30 real-valued features. The papers [29, 30, 28, 27] contain more detailed descriptions of this database. This data has been provided by The University of Wisconsin as the Wisconsin Diagnostic Breast Cancer Database. Later in the paper results are also presented for experiments with the Wisconsin Prognostic Breast Cancer Database.

The supervised assignment of elements of a given set into classes is the objective of classification analysis. Our approach is to find clusters of like points within the classes
that will accurately describe the classes. From the theoretical point of view, global optimisation methods provide the best solution for determining the centers of clusters within classes. In the simplest cases the well-known methods of convex optimisation can be used to allow one to solve for a single cluster. In this case the accuracy is high but in most instances it is necessary and desirable to deal with a larger number of clusters. In this paper we propose an algorithm for the solution of classification analysis problems, based on the methods of convex and global optimisation.

For the primary analysis of the given data we consider a special problem of convex programming. This problem allows us to:

- carry out feature selection - identify the variables, which are most important for the classification task and substantially reduce the dimension of the global optimisation problem,

- describe the classes by clusters - use global optimisation for finding the clusters to describe the classes.

We use the cutting angle method [2, 5, 21, 20] in global optimisation in order to solve this problem. The algorithm achieves improvement over the reported results for diagnosis and prognosis with these databases.

2 Approaches to Classification and Clustering

The classification problem involves assigning data examples into classes or clusters of like points. Statistical approaches [9] to this problem include cluster analysis [23] and multivariate analysis [11], [12]. Machine learning approaches can be found in [22, 23, 24] and integer and mathematical programming approaches in [6], [1], [15], [17], [19], [22]. Many approaches are based on a similarity measure. The nearest neighbour algorithm is often used as the basis of a similarity approach. He and Hawkins [10] compare the performance of this metric when the weights are determined from the mutual information measure, the usual Euclidean metric or optimised by using genetic algorithms or simulated annealing. Support Vector Machines (SVM) is a relatively new learning approach introduced by Vapnik in 1995 for solving two-class pattern recognition problems [25].

Techniques from artificial intelligence include self-organising maps, and minimum message length (MML) [26]. Bradley, Mangasarian, and Street [7] (see, also [15]) suggested the k-Median Algorithm. Here the problem of cluster analysis has been reduced to the problem of bilinear programming: minimising a bilinear function on a set defined by linear inequalities. This algorithm terminates in a finite number of iterations at a stationary point satisfying the minimum principle as a necessary condition for this mathematical programming problem. Trials of this algorithm with breast cancer prognosis provided very good results. Unfortunately, there is no guaranteed way to ensure global optimality of the solution obtained by either the k-Median or other known algorithms. So different starting points are usually used to initialize the algorithm ([15]) in order to examine the nature of the solutions. For example random starting cluster centers or some other heuristic can be used but even in this case the obtained solution is not necessarily a global one. It is also interesting to note that problems of cluster analysis are very close to the so-called location problems. An excellent review of data clustering can be found in [13].

Mangasarian [15] reports very good results in feature selection on classification problems, by attempting to use a piecewise linear approach to separating the data into classes. Many real problems are not piecewise linearly separable and the techniques employing convex and global optimisation can have advantages over this approach and other approaches that assume piecewise linearly separable sets.

The feature selection problem involves the selection of a subset of features that will be sufficient (and the most useful) in making predictions. Techniques for this from statistics include principal component analysis. Techniques from artificial intelligence include C4.5, a rule induction algorithm developed by Quinlan [18]. Mangasarian [16] illustrates a linear programming method called the multisurface method tree that reduced the number of features required for breast cancer diagnosis. The technique was used to reduce the number of features used in the diagnosis of breast cancer from 30 down to 3 and the result was one with 97.5% accuracy using cross validation. This accuracy rate is higher than any other diagnostic method except for surgical biopsy. However Mangasarian’s methods can not find the exact solution of the problem under consideration, so it can not be guaranteed that his methods will work well for other problems of this kind.

3 A global optimisation approach to classification

In this section we will describe the mathematical programming approach to classification.

Consider an \( n \)-dimensional space \( \mathbb{R}^n \). Elements of this space are \( n \)-dimensional vectors \( \mathbf{a} \), that is collections of \( n \) numbers: \( \mathbf{a} = (a_1, \ldots, a_n) \). Assume that \( \mathbb{R}^n \) is equipped with a norm \( \| \cdot \| \), then the distance between vectors \( \mathbf{a} \) and \( \mathbf{x} \) is \( \| \mathbf{a} - \mathbf{x} \| \). Different kinds of norm can be considered, in particular, the 1-norm, the 2-norm and the max-norm. By
definition, for a vector \( a = (a_1, \ldots, a^n) \) we have

\[
\|a\|_1 = \sum_{i=1}^{n} |a_i|, \quad \|a\|_2 = \sqrt{\sum_{i=1}^{n} (a_i)^2},
\]

\[
\|a\|_{\text{max}} = \max_{i=1, \ldots, n} |a_i|.
\]

Let \( A \) be a set of \( m \) \( n \)-dimensional vectors \( a_i = (a_i^1, \ldots, a_i^n), \quad i = 1, \ldots, m \). Assume that this set can be presented as the union of \( k \) clusters. Each cluster can be represented by a point, which can be considered as the centre of this cluster and a cluster is described by finding its centre. Thus we would like to find \( k \) points, which are centres of \( k \) clusters. Assume that we already have a set \( X \), consisting of \( k \) central points \( x_1, \ldots, x_k \). The distance \( d(a_i, X) \) from a point \( a_i \in A \) to this set is defined by

\[
d(a_i, X) = \min_{j=1, \ldots, k} \|a_i - x_j\|.
\]

The total distance \( d(A, X) \) from the set \( A \) to the set \( X \) is computed by the formula:

\[
d(A, X) = \sum_{i=1}^{m} d(a_i, X).
\]

The set \( \bar{X} \), which is a solution of the following mathematical programming problem:

\[
f_0(x_1, \ldots, x_k) := \sum_{i=1}^{m} \min_{j=1, \ldots, k} \|x_j - a_i\| \to \min,
\]

subject to \( x_j \in \mathbb{R}^n, \quad j = 1, 2, \ldots, k \)

can be considered as a set of centres of clusters, by assuming that the set \( A \) has \( k \) clusters. We can specify this problem by choosing a special norm (any of the 1-norm, the 2-norm or the max-norm). Note that the objective function \( f_0 \) in (1) is not necessarily differentiable.

If \( k = 1 \), then (1) has the form

\[
\sum_{i=1}^{m} \|x - a_i\| \to \min, \quad \text{subject to } x \in \mathbb{R}^n.
\]

This is a problem of (nonsmooth) convex programming and can be solved by various methods.

If \( k > 1 \) then the problem (1) becomes non-convex, hence we have a multi-extremal problem of non-smooth and non-convex optimisation. Various methods for the solution of this problem can be applied. We shall use the so-called cutting angle method [2, 5, 21, 20] for the solution of this problem.

The question of the number of clusters that will best describe a class arises and a priori we cannot determine the number, so we determine the number of clusters, step by step. We begin with the assumption that only one cluster gives a good description of the entire set \( A \). In this case we can consider the problem of unconstrained optimisation (2). This problem becomes a problem of nonsmooth convex programming where we need find only one vector, so the dimension of this problem is equal to \( n \). There are many methods for solving this problem and we used the method described in [3, 4]. When the problem is solved, it is useful to examine, whether the solution describes the entire set \( A \) well. This point will be discussed further in the next section for a special case.

Assume that one cluster does not give a good description of the set \( A \). In this case we seek the set of two clusters that better describes the set. The problem (1) becomes a very complicated problem of multi-extremal global optimisation and its dimension becomes \( 2n \) instead of \( n \). However, the centre of one cluster, which we have already obtained, contains important information on the set \( A \), so the question arises: is it possible to use this information in order to simplify the problem (1) and reduce its dimensionality. We shall address this question for a special, but very important case in the next section.

4 Dimensionality reduction

Global optimisation methods can provide the best solution for determining the centers of these clusters in our approach. Unfortunately both storage and time requirements in these methods are strongly dependent on the dimension of the problem. It is therefore important to give consideration to feature selection or dimensionality reduction methods. Consider the following situation: a set \( A \) of \( n \)-dimensional vectors is divided into \( p \) parts \( \mathcal{A}_1, \ldots, \mathcal{A}_p \).

Subsets \( \mathcal{A}_q \subset \mathcal{A}_q' \) \( (q = 1, \ldots, p) \) are known. We are interested in the following problem: for a given vector \( a \in A \), find a number \( q = 1, \ldots, p \) such that \( a \in \mathcal{A}_q \).

In the sequel we assume for the sake of simplicity that \( p = 2 \).

Assume that the set \( \mathcal{A}_q' \), for \( q = 1, 2 \) is the union of \( k_q \) clusters and let points \( x_q^j, \quad j = 1, \ldots, k_q, \quad q = 1, 2 \) be centres of these clusters. For each point \( a \in A \) we can find the nearest point \( x_q \) to \( a \) from the set of all centres of clusters for both sets \( \mathcal{A}_1, \mathcal{A}_2 \).

We accept the following hypothesis: Let \( q \) be a number, such that the point \( x_q \) is the centre of one of the clusters of the set \( \mathcal{A}_q \). Then \( a \in \mathcal{A}_q \).

Of course this hypothesis makes sense only if

1) the set \( \mathcal{A}_q \) is a good representation of the set \( \mathcal{A}_q' \), \( q = 1, 2 \);
2) centres of clusters of the set \( \mathcal{A}_q \) are good representatives of this set.
We suppose that 1) above is valid for the given sets $A_1$ and $A_2$ and shall concentrate only on item 2).

First let us assume that each set $A_1$, $A_2$ consists of one cluster. Using methods of convex programming we can find centres of these clusters. Denote them by $y_1$ and $y_2$, respectively. Compare the distances $||y_1 - a||$ and $||y_2 - a||$ for each point $a \in A_1 \cup A_2$. If $||y_1 - a|| < ||y_2 - a||$ for almost all $a \in A_1$, we can conclude that $y_1$ satisfactorily describes the set $A_1$. Otherwise we need to look for two clusters for this set. A similar conclusion can be made for the set $A_2$. Assume that at least one point $y_1$ does not satisfactorily describe the set $A_q$, $q = 1, 2$, then we need to look for new clusters. Now let us compare the coordinates of these points. If the $l$-th coordinates $y_{1l}$ and $y_{2l}$ of these points are close enough we can suggest that the coordinate $l$ does not contribute to distinguishing these two sets and not consider it. This suggestion allows us to substantially reduce the dimension of problem (1) and may be a useful heuristic. However, we need to elaborate further on what “close enough” means and this will be done later.

Assume that only coordinates from the set $L \subseteq \{1, \ldots, n\}$ remain, so instead of a vector $x = (x_l)_{l=1}^n$ we consider the vector $(x_l)_{l \in L}$. We shall denote this vector by the same symbol $x$. Let $|L|$ be the number of elements of $L$. Assume that $|L| < n$.

There are a number of alternative approaches for finding the two clusters (for the sake of definiteness we consider the set $A_1$).

1) We can forget about the known point $y_1$ and look for two new clusters, then the dimension of the problem (1) is $2|L|$.

2) We can assume that the point $y_1$ is the centre of one of the two clusters and look only for the centre of a second cluster. The dimension of the problem (1) in such a case is $|L|$.

3) Since the point $y_1$ is the centre of the “mixture” of two clusters, we can separate points into those close to $y_1$ and others and find for the close points, a new cluster by solving the convex programming problem, then repeat item 2).

4) We are looking for two clusters, which have (unknown) centres $y_{11}$ and $y_{12}$, respectively. Consider again the known point $y_1$ as the mixture of these two centres, then $y_1 = ty_{11} + (1-t)y_{12}$ with $0 \leq t \leq 1$, so we can seek only one of two unknown clusters and the number $t$. In this case the dimension of the problem (1) is $|L| + 1$.

5) We separate points from the first set, $A_1$, which are closer to $y_1$ than to $y_2$ and those that are closer to $y_2$. We consider only the points that are closer to $y_1$ for refining the centre $y_1$. Then we can repeat 2). The dimension of the problem is $|L|$.

We choose item 5) for the further consideration because it is of dimension $L$ and it incorporates information about the other set $A_2$ in the refinement process. Having two clusters we can continue in the same manner step by step to determine all clusters, that are needed.

## 5 The Algorithm with parameter selection

In this section the algorithm with the parameter selection will be described and then it will be applied to the Wisconsin Diagnostic Breast Cancer database and later the Wisconsin Prognostic Breast Cancer database.

We may consider the set $A_1$ as consisting of vectors related to benign cases and the set $A_2$ as consisting of vectors related to malignant cases. Let

$$N_1 = \{i | i = 1, \ldots, |A_1|\}$$

and

$$N_2 = \{i | i = |A_1| + 1, \ldots, |A_1| + |A_2|\}.$$

We now describe an algorithm for the solution of the problem of representing the classes by clusters. The first stages of the algorithm reduce the number of dimensions required.

### THE ALGORITHM

The first three steps of the algorithm constitute the feature selection phase. They describe the determination of the initial (single) clusters for each set $A_q$ and the selection of the most informative features. Step 4 is an evaluation step to test the feature reduction phase. Step 5 refines the centre of the cluster by removing the “misclassified” points before re-computing the cluster centres. Step 6 formulates a global optimisation problem to find an additional cluster to improve the description of the class by clusters. Figure 2 describes the situation indicating the refinement of the clusters by re-computing the centres $x^*_q$ to $x^0_q$ after removing the misclassified cases.

#### Step 1. Determination of centres of the clusters, by assuming that sets $A_1$ and $A_2$ contain a unique cluster.

Compute the centers of clusters solving the following problems of convex optimisation:

$$\sum_{i \in N_1} ||x_1 - a_i|| \to \min,$$

(3)

$$\sum_{i \in N_2} ||x_2 - a_i|| \to \min,$$

(4)
subject to \( x_j^l \geq 0, \ l = 1, \ldots, n, \ j = 1, 2 \).

**Step 2.** To find points of a set \( A_q \), \( q = 1, 2 \), which are closer to the centre of the cluster of the other set.

Let \( x_1^q \) and \( x_2^q \) be the solutions to the problems (3), (4), respectively. Compute the sets:

\[
N_1^q = \{ i \in N_1 | ||x_1^q - a_i|| \leq ||x_2^q - a_i|| \},
\]

\[
N_2^q = \{ i \in N_2 | ||x_1^q - a_i|| \leq ||x_2^q - a_i|| \}.
\]

These sets are essentially the points “misclassified” by the current clusters.

**Step 3.** To determine the informative parameters (coordinates).

Order the coordinates in the following way:

\[
| \langle x_1^q \rangle^1 - \langle x_2^q \rangle^1 \rangle \geq \ | \langle x_1^q \rangle^2 - \langle x_2^q \rangle^2 \rangle \geq \ldots \geq | \langle x_1^q \rangle^n - \langle x_2^q \rangle^n |.
\]

In a step by step fashion reduce the number of coordinates, beginning with the last one. On removing coordinates, recompute the sets \( N_1^q \) and \( N_2^q \) with respect to the rest of the vectors \( x_1^q, x_2^q \) and check the cardinality of these sets \( N_1^q \) and \( N_2^q \). In essence these are the misclassified cases and when the numbers in these sets shows a dramatic relative increase, the process terminates. Let \( L \) be a set of remaining coordinates.

**Step 4.** Evaluate the classification performance of the reduced feature set using the single cluster representations against the classification performance using the full feature set. The classification accuracy using the reduced feature set should not be significantly below that obtained when using all features.

In the next stage the dataset is split into training sets and test sets as described later.

**Step 5.** Refine the centre of a cluster by using only vectors, which are closer to the previous centre of a cluster.

Solve the following convex programming problems:

\[
\sum_{i \in N_1 \setminus N_1^q} ||x_1 - a_i|| \to \min, \tag{5}
\]

\[
\sum_{i \in N_2 \setminus N_2^q} ||x_2 - a_i|| \to \min, \tag{6}
\]

subject to \( x_j^l \geq 0, \ j = 1, 2, \ l \in L \).

(Here \( x_j \) and \( a_i \) are vectors of the reduced dimension: \( x_j = (x_j^l)_{l \in L}, a_i = (a_i^l)_{l \in L}. N_q \setminus N_q^* \) is the set of points that are close to the centre of \( A_q, \ q = 1, 2 \).)

**Step 6.** Let \( x_1^0 \) and \( x_2^0 \) be the solutions of the problems (5), (6), respectively. Determine the second cluster in addition to the currently improved cluster that will enhance the description of each class. This approach is taken to constrain the dimensionality of the global optimisation problems:

\[
\sum_{i \in N_1} \min \{ ||x_1 - a_i||, ||x_1^0 - a_i|| \} \to \min, \tag{7}
\]

\[
\sum_{i \in N_2} \min \{ ||x_2 - a_i||, ||x_2^0 - a_i|| \} \to \min. \tag{8}
\]

**Step 7.** Let \( x_1 \) and \( x_2 \) be the solutions of problems (7), (8), respectively. Evaluate the performance of the two cluster representation using the training data. Return to Step 5 if necessary and repeat Steps 5 and 6 if more clusters are necessary.

In Step 5 we now have two cluster centres for each set \( x_1^0, x_1 \), and \( x_2^0, x_2 \). There are many ways that we could...
proceed, but to be consistent with our earlier approach in Step 2 we define:

\[
N_1^0 = \{i \in N_1 \mid \|x_i^0 - a_i\| \leq \\
\min(\|x_i^0 - a_i\|, \|x_o^0 - a_i\|, \|x_j^0 - a_i\|)\}
\]

\[
N_1^1 = \{i \in N_1 \mid \|x_i^1 - a_i\| \leq \\
\min(\|x_i^1 - a_i\|, \|x_o^1 - a_i\|, \|x_j^1 - a_i\|)\}
\]

\[
N_2^0 = \{i \in N_2 \mid \|x_i^0 - a_i\| \leq \\
\min(\|x_i^0 - a_i\|, \|x_i^a - a_i\|, \|x_i^1 - a_i\|)\}
\]

\[
N_2^1 = \{i \in N_2 \mid \|x_i^1 - a_i\| \leq \\
\min(\|x_i^1 - a_i\|, \|x_i^a - a_i\|, \|x_i^1 - a_i\|)\}
\]

The convex programming problems in Step 5 now become:

\[
\sum_{i \in N_i^0} \|x_i - a_i\| \to \min,
\]

(9)

\[
\sum_{i \in N_i^1} \|x_i - a_i\| \to \min,
\]

(10)

\[
\sum_{i \in N_i^0} \|x_i - a_i\| \to \min,
\]

(11)

\[
\sum_{i \in N_i^1} \|x_i - a_i\| \to \min,
\]

(12)

subject to \(x_i^j \geq 0, j = 1, 2, l \in L\).

In the above we return to Step 5 to refine the cluster centres by convex optimisation techniques. This approach is taken rather than repeating Step 6 for the refinement to avoid the much heavier computation required in the global optimisation approach. In fact this refinement process has not been necessary in the experiments undertaken here and will be the subject of further work.

In Step 6 it would be more desirable to re-compute the new cluster centres as two new centres which optimally describe the class. However this problem is of higher dimensionality and requires further work which is currently underway.

6 Computational results for the diagnostic database

The Wisconsin Diagnostic Breast Cancer database consists of 569 vectors with known outcomes and it represents a training set with which a classifier can be constructed to diagnose future examples. It was created by W.H. Wolberg of the General Surgery Department, University of Wisconsin, Clinical Sciences Center, with W. N. Street and O.L. Mangasarian from the Computer Sciences Department, University of Wisconsin.

The first task in dealing with this data was to normalise the variables. This was done by scaling by a non-singular matrix \(T\) so that the values of all variables were between 0 and 10.

The set \(A_1\) consists of vectors related to benign cases and contains 357 vectors; the set \(A_2\) consists of vectors related to malignant cases and contains 212 vectors. So

\[N_1 = \{i \mid i = 1, \ldots, 357\}, \quad N_2 = \{i \mid i = 358, \ldots, 569\}\]

The feature selection phase of the algorithm described above is a heuristic that allows a significant reduction in the dimension of the problem which is very important for being able to use global optimisation techniques. It relies on using numerical methods of convex programming to form the initial clusters so that we are able to choose the most informative parameters. Our experience shows that for determination of clusters in the database under consideration it is sufficient to consider only three parameters: perimeter, area, and concavity. Of course, by taking more parameters we can more exactly describe the clusters, but our numerical experiments showed that the effect of the increase of the number of parameters is very small.

Using convex programming methods we established that all 30 parameters allow the determination of the set, which contains a given vector with 90% accuracy. In applying the feature selection phase of the algorithm there are only small changes in the number of misclassifications (cardinality of \(N_1^*\) and \(N_2^*\)) through to 5 parameters. The use of only 5 parameters: mean values of perimeter, area, compactness, concavity and concave points allows the determination of this set with almost 89% accuracy. The same result was obtained with the largest values (instead of the means) of the same parameters. For further investigations only the mean values of parameters were chosen. Following the algorithm then we considered the group of three parameters: mean values of perimeter, area and concavity. This group of parameters describes the data with 87% accuracy. For all pairs of parameters we obtained sharp increases in the cardinality of \(N_1^*\) and \(N_2^*\). In order to confirm that the algorithm calculated the best group of parameters we considered various groups of parameters which contained 4 or 3 parameters from the chosen 5. Various groups of 4 parameters describe the data with accuracy from 85% to 88%. The various groups of 3 parameters describe them with accuracy from 82% to 87%. The best result was obtained for the mean values of perimeter, area and concavity which coincides with the group of 3 parameters obtained by the algorithm. Taking into account the results of these experiments we can conclude that for computing the centres of clusters it is sufficient to use only the 3 parameters: mean values of...
By applying the algorithm and using only the mentioned parameters we computed two clusters for both sets $A_1$ and $A_2$ of vectors. We took 80% of all vectors as a training set and the remaining 20% of vectors were used for testing the clusters obtained on the basis of the training set. Five different variants of the training sets and test sets were considered. The results of the numerical experiments show that the Algorithm found two clusters for each set (2 for the malignant set and 2 for the benign set) which describe the remaining 20% of the vectors with 97 — 98% accuracy, (97 — 98.5% accuracy for benigns and with 96 — 100% accuracy for malignants.)

Similar experiments were carried out using the C4.5 algorithm to classify the test sets. The results based on using all 30 features produce an overall error rate of 6.7% (8.8% for the malignant set and 4.7% for the benign set).

In the numerical experiments we considered the 1-norm, the 2-norm and the max-norm. The results with using 1-norm and 2-norm were close enough. However, using the max-norm makes the objective function in the global optimisation problem very complicated and it has many points of global minimum. Using the maximum instead of the sum in the objective function also produces an objective function which is very complicated. The best results were obtained using the 1-norm, the 2-norm and the sum operation in the objective function.

It should be noted that in numerical experiments we also considered the cases with 4 and 5 parameters, but did not find considerable improvement in their results.

### 7 Computational results for the prognostic database

In this section the algorithm with the parameter selection will be applied to the Wisconsin Prognostic Breast Cancer database. This database consists of 198 vectors with 34 attributes. It was also created by W.H. Wolberg, W. N. Street and O.L. Mangasarian. In this database the second attribute is either disease-free time or recurrence time. On fields 4-33 ten real-valued features, their standard error and largest values are given. They are similar to Diagnostic database features. Finally, this database contains two traditional prognostic predictors: tumour size and the number of involved lymph nodes. The database consists of 151 nonrecurring and 47 recurrent cases. The papers [29, 30, 28, 27] contain more detailed descriptions of the database.

First we applied the algorithm in order to reduce the number of features. Our numerical experiments show that in order to obtain good description of the sets under consideration it is enough to take three parameters: perimeter, area and concavity. If concave points are used instead of concavity similar results are obtained. It should be noted that time is added as an input feature, so we have four features. Then we applied the algorithm for the calculation of centres of clusters for recurrence and non-recurrence cases separately. For the first set we calculated two clusters, whereas for second only one cluster. Thus, the centres of clusters are now in the space of computed features $\times$ time. This description provides an almost 100% accurate description of the sets. Then the centres of clusters are used to predict the time to recur and the disease-free time for a new patient. To achieve this we fix all features and only vary time. The time
Table 1. Average error for 2-norm and 1-norm

<table>
<thead>
<tr>
<th></th>
<th>All points</th>
<th>Non-recur</th>
<th>Recur</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-norm</td>
<td>12.8</td>
<td>13.1</td>
<td>12.5</td>
</tr>
<tr>
<td>1-norm</td>
<td>13.2</td>
<td>13.5</td>
<td>12.9</td>
</tr>
</tbody>
</table>

at which this point is closer to the centre of recurrence clusters than to the centre of non-recurrence cluster is the value of time for which this patient would go from being “non-recur” to being classified “recur”. Therefore, that time can be interpreted as a predicted time to recur.

The algorithm above was tested with leave-one-out testing to compute its accuracy in predicting future outcomes. The leave-one-out method is a special case of cross validation (see [14]). Here the predictive model is built using all but one of the examples and tested on the left-out case, this is repeated for each example, in turn, as the test case. In numerical experiments we used both Euclidean and 1-norm. The average error in time to recur (months) is shown in Table 1.

8 Conclusion

We have proposed an algorithm for the solution a broad class of classification problems based on the methods of convex and global optimisation using training sets. In this algorithm clusters are computed step by step with the starting cluster being computed by convex programming methods and all clusters being computed by using the Cutting Angle Method of global optimisation.

This algorithm has been applied to the Wisconsin Diagnostic and Prognostic Breast Cancer databases and indicates very good diagnostic and prognostic possibilities on the image features from the FNA nuclei. The numerical experiments reported above have demonstrated the effectiveness of the algorithm. The results achieved using this algorithm are at least as good as Mangasarian’s results and much better than C4.5. The computed cluster centers classify the data with very high accuracy. The global optimisation approach required the dimensionality of the problem to be restricted and the algorithm introduced here has been effective in reducing the number of parameters (features) significantly. Both of these problems, feature selection and clustering, are important in data mining. In this approach the most important features in the classification task have been discovered in quite an objective way without the application of expert knowledge. This algorithm with FNA and standard image processing techniques provides a means for achieving more accurate and effective medical diagnosis and prognosis for breast cancer patients.

Further work on developing the algorithm is in progress. Work with other databases also from the medical domain is being carried out as well as comparisons with other more well known statistical techniques.

References


