Journal of Computer Science 6 (3): 369-373, 2010
ISSN 1549-3636

A Derivative-Free Optimization Method for Solving Classification Problem

Parvaneh Shabanzadeh, Malik Abu Hassan and Wah June Leong
Department of Mathematics, Faculty of Science,
University Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Abstract: Problem statement: The aim of data classification is to establish rules for the classification of some observations assuming that we have a database, which includes of at least two classes. There is a training set for each class. Those problems occur in a wide range of human activity. One of the most promising ways to data classification is based on methods of mathematical optimization. **Approach:** The problem of data classification was studied as a problem of global, nonsmooth and nonconvex optimization; this approach consists of describing clusters for the given training sets. The data vectors are assigned to the closest cluster and correspondingly to the set, which contains this cluster and an algorithm based on a derivative-free method is applied to the solution of this problem. **Results:** Proposed method had been tested on real-world datasets. Results of numerical experiments had been presented which demonstrate the effectiveness of the proposed algorithm. **Conclusion:** In this study we had studied a derivative-free optimization approach to the classification. For optimization generalized pattern search method has been applied. The results of numerical experiments allowed us to say the proposed algorithms are effective for solving classification problems at least for databases considered in this study.

Key words: Classification, direct search, nonsmooth optimization

INTRODUCTION

The aim of data classification is to establish rules for the classification of some observations assuming that the classes of data are known. To find these rules, a researcher can use known training subsets of the specified classes. The construction of a classification procedure may also be a pattern recognition procedure, a discrimination procedure or supervised learning procedure. Those problems occur in a wide range of human activity.

Many methods exist for data classification, which are based on quite different approaches (neural networks, statistics, and methods of information theory). Michie *et al*. (1994) explains an excellent review of these methods, including their computational investigation and comparison.

One of the most promising ways to data classification is based on methods of mathematical optimization. For supervised classification we have a database, which includes of at least two classes. There is a training set for each class and there are two different ways for the application of optimization. The first, which we shall call outer, is based on the separation of the given training sets by means of a certain function. The outer approach is currently the most popular, for example by Mangasarian (1997) and Bradley and Mangasarian (2000), where problems of quadratic and bilinear programming are applied for classification and then linear programming techniques are used for the solution of these problems. The second (inner) approach consists of describing clusters for the given training sets. The data vectors are allocated to the closest cluster and correspondingly to the set, which includes this cluster. The conceptual description of this approach can be found in Bagirov *et al*. (2001). Numerical experiments show that for supervised classification of databases of a small to medium size, the inner approach presents a more precise description of databases than the outer approach. We apply the inner approach in this study. For the execution of this approach one needs to solve a complex problem of non-convex and non-smooth unconstrained optimization. Global methods provide more precise descriptions of clusters. A powerful method for solving nonsmooth optimization problems (the generalized pattern search method GPS) has been developed (Torczon, 1997; Audet and Dennis, 2003). Too, the numerical experiments is presented in this study which show that the inner approach to the supervised classification

Corresponding Author: Parvaneh Shabanzadeh, Department of Mathematics, Faculty of Science, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia Tel: 0060173443492 Fax: 0060389466834

problem based on optimization techniques gives results close to the best known method.

MATERIALS AND METHODS

The global optimization algorithm to classification: At the first we introduce a formulation to the classification problem in terms of global optimization.

Consider the dataset which contains k classes, that is, k nonempty finite subsets B_j, $j=1,...,k$ of m-dimensional space R^m. Assume that the set B_j consists of d_j points ($j=1,...,k$). The task of classification is to establish means where by we can categorize a new observation into one of the existing classes. Therefore in order to solve this problem we suggest finding clusters for each set B_j, $j=1,...,k$ and identifying these sets with the centers of the corresponding clusters. New observations are allocated to the class with least distance between its centers and these observations.

First, we will find the clusters of a finite set. Many approaches exist for solving this problem. We suggest a method based on global optimization ways mentioned in (Bagirov *et al*., 2001). Numerical experiments verify that this method outperforms known ones for many real-world databases.

Consider a set B which consists of d m-dimensional vectors $b^i = b^i_1,...,b^1_m, i=1,...,d.$ Assume that this set can be presented as the union of p clusters. Suppose also, that each cluster can be presented by a point, which can be considered as the center of this cluster. For finding a cluster we should find its center. Thus we would like to find p points which are centers of clusters. Thus the cluster analysis problem can be shown by the following problem of mathematical programming:

Minimize: $f(x^1,...,x^p)$

Subject to: $(x^1,...,x^p)\in R^{m\times p}$

Where:

$$f(x^1,...,x^k)=\sum_{i=1}^{d}\min_{s=1,...,p}\left\|x^s-b^i\right\| \quad (1)$$

Recall that $\|x\|_q=(\sum_{t=1}^{m}|x_t|^q)^{1/q}, 1\le q\le+\infty$

Note if p>1, then the objective function f in the problem (1) is nonsmooth and nonconvex. We call f the cluster function.

Some problems happen when the proposed procedure is applied. Note that the number of variables in the global optimization problem (1) is $p \times m$. If the number p of clusters and the number m of attributes are large, then we have a large-scale global optimization problem. On the other hand it is difficult to define, a priori how many clusters represent the set B under consideration. Therefore we need to consider different numbers of clusters, starting from a certain small number p. If the solution of the corresponding optimization problem (1) is not satisfactory, we need to consider the problem (1) with p + 1 clusters and so on. Thus we need to solve repeatedly the arising global optimization problem (1) with different p.

Therefore assuming that the set B consists of only one cluster we can calculate its center by solving the following convex programming problem:

Minimize: $f_1(x)=\sum_{i=1}^{d}\left\|x-b^i\right\|$

Subject to: $x\in R^m$ (2)

Removing all misclassified points and solving problem (2) again we create this center more precise. We will indicate this center by x^1. In order to find a center of the second cluster we solve the following problem of global optimization:

Minimize: $f_2(x)=\sum_{i=1}^{d}\min\left\{\left\|x^1-b^i\right\|,\left\|x-b^i\right\|\right\}$

Subject to: $x\in R^m$ (3)

Suppose that we have already calculated the center x^{t-1} of $(t-1)$-th cluster, then the center x^t of t-th cluster is described as a solution to the following problem:

Minimize: $f_t(x)=\sum_{i=1}^{d}\min\left\{\left\|x^1-b^i\right\|,\left\|x^2-b^i\right\|,...,\left\|x^t-b^i\right\|,\left\|x-b^i\right\|\right\}$

Subject to: $x\in R^m$ (4)

Then the number of variables in (4), which is m, is significantly less than that in (1).

Remark 1: It is possible that the number of clusters calculated step-by-step is greater than the number of clusters, which can be found directly by solving (1). However, even in such a case the solution of (1) needs much more time than the solution of the series of problems (4).

The algorithm for classification: In continue we give a description of the algorithm for the solution of classification problems.

We consider a database which contains 2 classes: B_1 and B_2. Let:

$$P_1 = \{1,...,|B_1|\}$$
$$P_2 = \{|B_1|+1,...,|B_1|+|B_2|\}$$

Let $\varepsilon>0$ be a tolerance.

Algorithm 1: Classification algorithm:

Step 1: Initialization. Determine centers of clusters, by assuming that sets B_1 and B_2 contain a unique cluster. Compute the centers of clusters solving the following problems of convex optimization:

$$\text{Minimize: } \sum_{i\in P_1}\left\|x^1 - b^i\right\| \quad (5)$$

$$\text{Minimize: } \sum_{i\in P_2}\left\|x^2 - b^i\right\| \quad (6)$$

Subject to: $x^j \in R^m, j=1,2$

Set r = 1. Let x_{1r}^* and x_{2r}^* be the solutions to the problems (5) and (6) and allow f_{1r}^* and f_{2r}^* be the values of these problems, respectively.

Step 2: Compute the sets:

$$P_{1r}^* = \left\{i\in P_1 : \min_{t=1,...,r}\left\|x_{2t}^* - b^i\right\| \le \min_{t=1,...,r}\left\|x_{1t}^* - b^i\right\|\right\}$$
$$P_{2r}^* = \left\{i\in P_2 : \min_{t=1,...,r}\left\|x_{1t}^* - b^i\right\| \le \min_{t=1,...,r}\left\|x_{2t}^* - b^i\right\|\right\}$$

For find the sets of points "misclassified" by the current clusters.

Step 3: Compute the following sets:

$$K_1 = \left\{i\in P_1 \setminus P_{1r}^* : \left\|x_{1r}^* - b^i\right\| \le \min_{t=1,...,r-1}\left\|x_{1t}^* - b^i\right\|\right\}$$
$$K_2 = \left\{i\in P_2 \setminus P_{2r}^* : \left\|x_{2r}^* - b^i\right\| \le \min_{t=1,...,r-1}\left\|x_{2t}^* - b^i\right\|\right\}$$

Step 4: Improve the center of the cluster by solving the following convex programming problems:

$$\text{Minimize: } \sum_{i\in K_1}\left\|x^1 - b^i\right\| \quad (7)$$

$$\text{Minimize: } \sum_{i\in K_2}\left\|x^2 - b^i\right\| \quad (8)$$

Subject to: $x^j \in R^m, j=1,2$

Allow x^{01} and x^{02} be the solutions of the problems (7) and (8), respectively. Set $x_{1r}^* = x^{01}$ and $x_{2r}^* = x^{02}$.

Step 5: Determine the next cluster. Solve the following optimization problems:

$$\text{Minimize: } \sum_{i\in P_1}\left\{\left\|x^1 - b^i\right\|, \left\|x_{11}^* - b^i\right\|,...,\left\|x_{1r}^* - b^i\right\|\right\} \quad (9)$$

$$\text{Minimize: } \sum_{i\in P_2}\left\{\left\|x^2 - b^i\right\|, \left\|x_{21}^* - b^i\right\|,...,\left\|x_{2r}^* - b^i\right\|\right\} \quad (10)$$

Subject to: $x^j \in R^m, j=1,2$

Step 6: Allow x^{11} and x^{12} be the solutions and $f_{1,r+1}$ and $f_{2,r+1}$ be the values of the problems (9) and (10), respectively. Set $x_{1,r+1}^* = x^{11}$ and $x_{2,r+1}^* = x^{12}$.

Step 7: Checking the stopping criterion.

If:

$$\max\left\{\frac{|f_{1,r+1} - f_{1r}|}{f_{11}}, \frac{|f_{2,r+1} - f_{2r}|}{f_{21}}\right\} < \in$$

then the algorithm ends. Otherwise set k = k + 1 and go to Step 2.

Remark 2: In order to apply this algorithm to the investigation of concrete datasets we need to solve the minimization problem (4), since both (9) and (10) have the form (4).

Method for a global optimization In this part of paper we will discuss an algorithm for solving problems (9) and (10) in the classification algorithm. Since these functions are nonsmooth and evaluation of subgradients is difficult, direct search methods of optimization seem to be the best option for solving problems.

The main attraction of direct search methods is their ability to find optimal solutions without the need for computing derivatives, in contrast to the more familiar gradient-based methods. Among such methods the Generalized Patterns Search (GPS) methods which are well suited for the nonsmooth optimization.

The original pattern search methods are designed by Hooke and Jeeves (1961) for unconstrained optimization. Owing to their simplicity and practical use, pattern search methods have been still widely used. Recently, many researchers paid attention to pattern search methods for unconstrained optimization and did a lot of work on them, including Dennis and Torzon (1991), Generalized Pattern Search method (GPS) of Torczon (1997), Audet and Dennis (2003) and Coope and Price (2001). An interesting characteristic of the pattern search method is that it is simple and easy to implement and it only needs the ability to evaluate the function at a point.

GPS method: Consider the following problem:

Minimize: f(a)

where, $a \in R^n, f : R^n \to R(R^n$ shows the n-dimensional real search space).

We describe a generating set (positive spanning set) D as a set of vectors whose non-negative linear combinations span R^n. For example, a positive spanning set D for R^n could be $\{e_1, e_2, ..., e_n, -e_1, -e_2, ..., -e_n\}$, e_i is the i-th unit Cartesian vector in R^n. We mention that this set must contain at least n + 1 vectors to guarantee non-negative linear combinations and hence need not be unique. This method will take steps through comparing function values at each of the points defined by one of the search directions and step lengths. We will suppose Δ_h be the step length control parameter and let Δ_{tol} be the tolerance used to test for convergence.

Suppose that the algorithm starts with an initial guess a_0 that has a finite function value and an initial step length Δ_0. Then the GPS method can be described as follows:

Algorithm 2: Generalized pattern search:

1: Select generating set D(for example, let $D = \{e_1, e_2, ..., e_3, -e_1, -e_2, ..., -e_n\}$)
2: Choose Δ_0
3: for h = 1, 2, … do
4: if there exists $d_h \in D$ such that $f(a_h + \Delta_h d_h) < f(a_h)$ then
5: Set $a_{h+1} = a_h + \Delta_h d_h$ ▷ update the iterate
6: Set $\Delta_{h+1} = \Delta_h$ ▷ no change to the step length control parameter
7: else if
8: Set $a_{h+1} = a_h \triangleright f(a_h + \Delta_h d_h) \geq f(a_h)$ for all $d_h \in D$; do not update the iterate
9: Set $\Delta_{h+1} = \frac{1}{2}\Delta_h \triangleright$ contract the step length control parameter
10: if $\Delta_{h+1} < \Delta_{tol}$ then
11: GPS algorithm has converged
12: end if
13: end if
14: end for

Steps of the GPS algorithm can be generalized further; for instance, in step 1 the lengths of the vectors in the generating set can take on any values between specified lower and upper bounds; also, a finite number of additional search directions (other than the ones already included in the generating set), may be increased using physics based approach or any heuristics that seems suitable; for example, Latin hypercube search, random search, or a few generations of a genetic algorithm. This adds an optional search step in the each iteration of the GPS algorithm. The search through the directions of the generating set is commonly referred to as a local poll step. In step 4, the function value may require a large decrease. Finally, various scale factors may be used to update the step length control parameter Δ_h; therefore it is not always 1 in step 6 and 1/2 in step 9. These generalizations allow great freedom in using the GPS method and can have an important influence on the efficiency of the algorithm.

RESULTS

In this study we present results of the numerical experiments. The proposed algorithms have been tested on real-world datasets. The diabetes dataset, the liver-disorder dataset and the heart disease dataset have been used in numerical experiments. The explanation of these datasets can be found in Murphy and Aha (1991).

Remark 3: The number of iterations evaluated by the GPS method in Algorithm 1 is restricted. The cutting angle method evaluates at most 88 iterations for all cases. Stopping criterion $\varepsilon = 10^{-2}$ is used for Step 7 of Algorithm 1. In numerical experiments we use Algorithm 1 with the global optimization in Step 5. Algorithm 2 is used for global optimization and q-norms with q = 1 and q = 2. For the comparison of the results of numerical experiments we choose the algorithm of classification from Bradley and Mangasarian (1998) obtained by support vector machines SVM algorithm and results of numerical experiments using this algorithm.

Table 1: Results for real-world database

				q = 1			q = 2			SVM	
Data set	m	f	c	etr	ets	n	etr	ets	n	etr	ets
Heart	297	13	2	0.132	0.203	8	0.121	0.192	9	0.153	0.241
Liver	345	6	2	0.292	0.382	14	0.350	0.352	15	0.398	0.390
Diabetes	768	8	2	0.263	0.210	8	0.241	0.21	8	0.240	0.250

The code has been written in Matlab and the numerical experiments have been carried out on a PC Intel(R) Pentium(R) Dual with CPU 997 MHz. The results of numerical experiments are presented in Table 1.

DISCUSSION

For first database, "heart database come from the Cleveland Clinic Foundation and it is part of the collection of databases at the University of California. The liver-disorder database was donated by Richard S. Forsyth BUPA Medical research Ltd. The diabetes database, this database was originally given by Vincent Sigillito, Applied Physics Laboratory, John Hopkins University, Laurel, USA and was constructed by constrained selection from a larger database held by the National Institute of Diabetes and Digestive and Kidney Diseases (Bagirov *et al*., 2002)". The results presented in Table 1 show that the accuracies of our method for all of database with both norm are almost the same and also they are high enough and same or better accuracy than SVM.

CONCLUSION

In this study we have studied a global optimization approach to the classification. Classes in the database mentioned earlier are considered by using cluster centers in these classes so that for each class, the cluster analysis problem is solved. The last problem is studied as an optimization problem with nonconvex and nonsmooth objective function. Optimization is carried out by generalized pattern search. Numerical experiments using real-world databases have been carried out in order to verify the effectiveness of the proposed approach. The described method is effective for solving classification problems at least for databases studied in this study. It is interesting to consider methods are explained in this study with databases which contain more than two classes and a large number of observations and their numerical analysis will be theme of our future studies.

REFERENCES

Audet, C. and J.E. Dennis, 2003. Analysis of generalized pattern searches. Siam J. Optimiz., 13: 889-903. DOI: 10.1137/S1052623400378742

Bagirov, A.M., A.M. Rubinov and J. Yearwood, 2001. Using global optimization to improve classification for medical diagnosis and prognosis. Top. Health Inform. Manage., 22: 65-74. PMID: 11680278

Bagirov, A.M., A.M. Rubinov and J. Yearwood, 2002. A global optimization approach to classification. Optimiz. Eng., 3: 129-155. DOI: 10.1023/A:1020911318981

Bradley, P.S. and O.L. Mangasarian, 1998. Feature selection via concave minimization and support vector machine. http://reference.kfupm.edu.sa/content/f/e/feature_selection_via_concave_minimizati_74808.pdf

Bradley, P.S. and O.L., Mangasarian, 2000. Massive data discriminate via linear support vector machines. Optimiz. Methods Software, 13: 1-10. DOI: 10.1080/10556780008805771

Coope, I.D. and C.J. Price, 2001. On the convergence of grid-based methods for unconstrained minimization. Siam J. Optimiz., 11: 859-869. DOI: 10.1137/S1052623499354989

Dennis, J.E. and V. Torczon, 1991. Direct search methods on parallel machines. Siam J., Optimiz, 1: 448-474. DOI: 10.1137/0801027

Hooke, R. and T.A. Jeeves, 1961. Direct search solution of numerical and statistical problems. J. Assoc. Comput. Mach., 8: 212-219. DOI: 10.1023/A:1013760716801

Mangasarian, O.L., 1997. Mathematical programming in data mining. Data Min. Knowl. Discov., 1: 183-201. DOI: 10.1023/A:1009735908398

Michie, D.D., J. Spiegelhalter and C.C. Taylor, 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood Series in Artificial Intelligents, London. http://www.amsta.leeds.ac.uk/~charles/statlog/

Murphy, P.M. and D.W. Aha, 1991. UCI repository of machine learning databases. Technical report. http://archive.ics.uci.edu/ml/datasets.html

Torczon, V., 1997. On the convergence of pattern search algorithms. Siam J. Optimiz., 7: 1-25. DOI: 10.1137/S1052623493250780