New Optimization Methods in Data Mining

S. Özöğür-Akyüz,[1] B. Akteke- Öztürk[1], T. Tchemisova,[2] and G.-W. Weber[1]

[1] Institute of Applied Mathematics, METU, Turkey,
sozogur@metu.edu.tr, boztur@metu.edu.tr, gweber@metu.edu.tr
[2] Department of Mathematics, University of Aveiro, Portugal
tatiana@ua.pt

Summary. Data mining is a modern area of science dealing with the learning from given data in order to make predictions and estimations. Applications of Data mining can be found in various areas of academical and non academical life. This paper introduces new contributions by continuous optimization as a key technology in data mining. The methods suggested for solution of such important problems as clustering and classification, were recently obtained by the authors in collaboration with members of EURO working group EUROPT.

Key words: Data Mining, Classification, Clustering, Optimization, Statistic Learning, Infinite Programming

1 Introduction

Generally speaking, an optimization problem consists in maximization or minimization of some function (objective function) $f : S \to \mathbf{R}$. The *feasible* set $S \subseteq \mathbf{R}^n$ can be either finite or infinite, and can be described with the help of a finite or infinite number of equalities and inequalities or in the form of some topological structure in $\mathbf{R}^n$. The methods for solution of certain optimization problem depend mainly on the properties of the objective function and the feasible set. Thus, when we look for extrema of a linear function regarded on some polyhedral set, then the methods of *linear programming* can be applied; when f is a convex function and S is a convex set, we apply methods of *convex programming*; if the feasible set S is defined by infinite number of equalities or inequalities, the methods of *semi-infinite programming* should be used, etc. In this paper, we discuss how specific optimization methods of optimization can be used in some specific areas of data mining, namely, in *classification* and *clustering* that are considered interrelated [11].

2 Clustering

Clustering is a unsupervised learning in which data are separated into clusters according to their similarity. It has many applications, including decision-making and machine-learning, information retrieval and

medicine, image segmentation and pattern classification, etc. Alternatively, it may support preprocessing steps for other algorithms, such as classification and characterization, operating the detecting clusters [6].

2.1 Optimization models for clustering problems

Assume that we have a finite set X of points (patterns) in the $n-$dimensional space $\mathbf{R}^n : X = \{x^1, x^2, \ldots, x^M\}$, where $x^k \in \mathbf{R}^n$ $(k = 1, 2, \ldots, M)$. Given a number $q \in \mathbf{N}$, we are looking for q subsets C^i, $i = 1, 2, \ldots, q$, such that the medium distance between the elements in each subset is minimal and the following conditions are satisfied: 1. $C^i \neq \emptyset$, $(i = 1, 2, \ldots, q)$, 2. $X = \bigcup_{i=1}^{q} C^i$. As a measure of similarity we use any distance function. Here for the sake of simplicity we consider Euclidean distance $||\cdot||_2$. The sets C^i $(i = 1, 2, \ldots, q)$, introduced above are called *clusters* and the problem of determination of clusters is the *clustering problem*. When the clusters can overlap, the clustering problem is *fuzzy*. If we request additionally: 3. $C^i \bigcap C^j = \emptyset$ if $i \neq j$, then we obtain a *hard* clustering problem. Let us assume that each cluster C^i, can be identified by its *center* or *centroid*, defined as (see [3]) $c^i := \frac{1}{|C^i|} \sum_{x \in C^i} x$, where $|C^i|$ denotes a cardinality of the cluster C^i. Then the clustering problem can be reduced to the following optimization problem, which is known as a *minimum sum of squares clustering* [4]:

$$\begin{array}{l} \min \ \frac{1}{M} \sum_{i=1}^{q} \sum_{x \in C^i} ||c^i - x||_2^2 \\ \text{such that } \ C = \{C^1, C^2, \ldots, C^q\} \in \bar{C}, \end{array} \tag{1}$$

where $\bar{C}$ is a set of all possible $q-$partitions of the set X.

The clustering problem (1) can be rewritten as single *mixed-integer* minimization problem as follows:

$$\begin{array}{l} \min \ \ \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{q} w_{ij} ||x^j - c^i||_2^2, \\ \text{such that } \ \ w_{ij} \in \{0, 1\}, \ \ \sum_{i=1}^{q} w_{ij} = 1 \\ (i = 1, 2, \ldots, q) \ \ (j = 1, 2, \ldots, M). \end{array} \tag{2}$$

Here, centroids are rewritten as $c^i := (\sum_{j=1}^{M} w_{ij} x^j)/(\sum_{j=1}^{M} w_{ij})$, w_{ij} is the association weight of the pattern x^j with cluster i given by

$$w_{ij} = \begin{cases} 1, & \text{if pattern } j \text{ is allocated to cluster } i, \\ 0, & \text{otherwise.} \end{cases}$$

It can be shown that (2) is a global optimization problem with possibly many local minima [3]. In general, solving the global optimization problem is a difficult task. This makes it necessary to develop clustering algorithms which compute the local minimizers of problem (2) separately. In [3], the optimization techniques are suggested that are based on nonsmooth optimization approach. Finally, note that the clustering problems (1) and (2) can be reformulated as an unconstrained non smooth and non convex problem

$$\min f(c^1, c^2, \ldots, c^q), \tag{3}$$

where $f(c^1, c^2, \ldots, c^q) = \frac{1}{M} \sum_{i=1}^{M} \min_j ||c^j - x^j||_2^2$. Since the function $\psi(y) = ||y - c||_2^2$ $(y \in \mathbf{R}^n)$, is separable (as a sum of squares),

the function $\varphi(x^i) = \min_j ||c^j - x^j||_2^2$ is *piece-wise separable*. It is proved in [2] that the function $f(c^1, c^2, \ldots, c^q)$ is piecewise separable as well. The special separable structure of this problem together with its non smoothness allows a corresponding analysis and specific numerical methods related with *derivative free optimization*.

2.2 Cluster stability using minimal spanning trees

Estimation of the appropriate number q of clusters is a fundamental problem in cluster analysis. Many approaches to this problem exploit the *within-cluster dispersion matrix* (defined according to the pattern of a covariance matrix). The span of this matrix (column space) usually decreases as the number of groups rises and may have a point in which it "falls". Such an "elbow" on the graph locates in several known methods, a "true" number of clusters. Stability based approaches, for the cluster validation problem, evaluate the partition's variability under repeated applications of a clustering algorithm on samples. Low variability is understood as high consistency of the results obtained and the number of clusters that minimizes cluster stability is accepted as an estimate for the "true" number of clusters. In [10], a statistical method for the study of cluster stability is proposed. This method suggests a geometrical stability of a partition drawing samples from the partition and estimating the clusters by means of each one of the drawn samples. A pair of partitions is considered to be consistent if the obtained divisions match. The matching is measured by a *minimal spanning tree (MST)* constructed for each one of the clusters and the number of edges connecting points from different samples is calculated. MSTs are important for several reasons: they can be quickly and easily computed with the help known methods of *discrete* optimization (Prim's, Kruskal's or Dijkstra's algorithms, for example), they create a sparse subgraph which reflects some essence of the given graph, and they provide a way to identify clusters in point sets.

3 Classification in statistical learning

The problems of supervised data classification arise in many areas including management science, medicine, chemistry etc. The aim of *supervised learning* is to establish rules for the classification of some observations assuming that the classes of data are known. Classification is a supervised learning in which the classification function is determined from the set of examples so called training set.

3.1 Classification by SVM

In this paper, we concentrate on *support vector machines (SVMs)* as one important classification tool that uses continuous optimization [7]. A SVM is a classification method based on finding a discriminative function which maximizes the distance between two class of points. More formally, let (x, y) be an (input,output) pair, where $x \in \mathbf{R}^n$ and $y \in \{-1, 1\}$ and x comes from some input domain X and similarly y

comes from some output domain Y. A training set is defined by l input-output pairs by $S = \{(x_i, y_i)\}_{i=1}^{l}$. Given S and a set of functions $\mathcal{F}$ we search for a candidate function $f \in \mathcal{F}$ such that $f : x \mapsto y$. We refer to this candidate function as a *hypothesis* [5]. The classes are separated by an affine function, hyperplane $\langle w, x\rangle + b = 0$, where $w \in \mathbf{R}^n$ is a normal vector (weight vector) helping to define the hyperplane, $b \in \mathbf{R}$ is the bias term [5], and $\langle \cdot, \cdot \rangle$ denotes the scalar product. Hence, given a set of examples S, the SVM separates it into two groups by a hyperplane. In linearly inseparable cases, one can define a *non-linear mapping* ϕ which transforms the input space into a higher dimensional *feature space* that that we will refer to as the SVM. The original points are separable in feature space. But the mapping can be of very high-dimension or even infinite. Hence, it is hard to interpret decision (classification) functions which are expressed as $f(x) = \langle w, \phi(x)\rangle + b$. Following the notation of [5], the *kernel function* is defined as an inner product of two points under the mapping ϕ, i.e., $\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle$ which can also be explained as the similarity between two points and the optimization problem for separating two classes is expressed as follows:

$$\begin{array}{c} \min_{\xi,\mathbf{w},b} \|\mathbf{w}\|_2^2 + \mathcal{C}\sum_i \xi_i \\ \text{such that} \quad y_i \cdot (\langle \mathbf{w}, \phi(\mathbf{x_i})\rangle + b) \geq 1 - \xi_i \quad (i = 1, 2, \ldots, m), \end{array} \tag{4}$$

where $\mathcal{C}$ is an error constant to penalize tolerance variable, slack variable, ξ. The dual problem in the soft margin case looks as follows:

$$\begin{array}{c} \max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m} y_i y_j \alpha_i \alpha_j \kappa(x_i, x_j), \\ \text{such that} \quad \sum_{i=1}^{m} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq \mathcal{C} \quad (i = 1, 2, \ldots, m), \end{array} \tag{5}$$

where the vector α is dual variable so called *support vectors*. The solution of the optimization problem (5) yields a maximal margin hyperplane that defines our SVM. In [8], we propose a combination of infinitely many kernels in Riemann Stieltjes integral form for binary classification to allow all possible choices of kernels into the kernel space which makes the problem infinite in both dimension and number of constraints, a so called *infinite programming (IP)*. Based on motivation in [9], we can define our *infinite* learning problem as follows:

$$\begin{array}{l} \max\limits_{\theta \in \mathbf{R}, \beta} \ \theta \ (\beta : [a, b] \to \mathrm{R} \text{ monotonically increases}), \\ \text{such that } \int_{\Omega} \left(\frac{1}{2} S(\omega, \alpha) - \sum_{i=1}^{l} \alpha_i\right) d\beta(\omega) \geq \theta \ \forall \alpha \in A, \\ \int_{\Omega} d\beta(\omega) = 1. \end{array} \tag{6}$$

Here, $S(\omega, \alpha) := \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j, \omega)$, $A := \{\alpha \in \mathbf{R}^l \mid 0 \leq \alpha_i \leq \mathcal{C} \ \ (i = 1, 2, \ldots, l)$, and $\sum_{i=1}^{l} \alpha_i y_i = 0\}$. Let $T(\omega, \alpha) := S(\omega, \alpha) - \sum_{i=1}^{l} \alpha_i$, and $\Omega := [0, 1]$. Having introduced Riemann-Stieltjes integrals via functions β, we can now reinterpret the latter ones by (probability) measures. Herewith, (6) turns into the following form:

$$\begin{array}{l} \max\limits_{\theta \in \mathbf{R}, \beta} \ \theta \ (\beta : \text{ a positive measure on } \Omega), \\ \text{such that} \quad \theta - \int_{\Omega} T(\omega, \alpha) d\beta(\omega) \leq 0 \ \forall \alpha \in A, \ \int_{\Omega} d\beta = 1. \end{array} \tag{7}$$

It is evident [1] that problem (7) is an *infinite programming (IP) problem*. The dual to (7) is:

$$\begin{array}{l} \min\limits_{\sigma \in \mathbf{R}, \rho} \ \sigma \quad (\rho : \text{ a positive measure on} A), \\ \text{such that} \quad \sigma - \int_{A} T(\omega, \alpha) d\rho(\alpha) \leq 0, \ \forall \omega \in \Omega, \ \ \int_{A} d\rho(\alpha) = 1. \end{array} \tag{8}$$

Assume that there exist pairs $(\beta,\ \theta)$ and $(\rho,\ \sigma)$ of feasible solutions of problems (7) and (8) which are complementary slack, i.e., $\sigma^* = \int_A T(\omega,\alpha)d\rho^*(\alpha)$ and $\theta^* = \int_A T(\omega,\alpha)d\beta^*(\omega)$. Then, β has measure only where $\sigma = \int_A T(\omega,\alpha)d\rho(\alpha)$ and ρ has measure only where $\theta = \int_\Omega T(\omega,\alpha)d\beta(\omega)$ which implies that both solutions are optimal for their respective problems. The regularity condition of problem (8) is analyzed in [8]. The so-called *reduction ansatz* enables the Implicit Function Theorem for reducing an infinite number of constraints to a finite number [14]. Of course, this can also be achieved by a smart *discretization*. Note that we can also focus on parametric classes of probability measures; then our IP problems turn to *SIP* (semi-infinite programming) problems; eventually, when applying any of the those three approaches, we arrive at a finitely constrained program.

3.2 Max-min separability

According to [2], the problem of supervised data classification can be reduced to a number of set separation problems. For each class, the training points belonging to it have to be separated from the other training points using a certain, not necessarily linear, function. This problem is formulated in [2] as a nonsmooth optimization problem with max-min objective function. Let A and B be given disjoint sets containing m and p vectors from $\mathbf{R}^n$, respectively: $A = \{a^1,\ldots,a^m\},\ B = \{b^1,\ldots,b^p\}$. Let $H = \{h_1,\ldots,h_l\}$ be a finite set of hyperplanes, where h_j is given by $\langle x_j,z\rangle - y_j = 0\ j = (1,2,\ldots,l)$ with $x_j \in \mathbf{R}^n,\ y_j \in \mathbf{R}$. Let $J = \{1,2,\ldots,l\}$. Consider any partition of J in the form $J^r = \{J_1,\ldots,J_r\}$, where $J_k \neq \emptyset, k = 1,\ldots,r;\ J_k \bigcap J_s = \emptyset$, if $k \neq s;\ \bigcup_{k=1}^r J_k = J$. Let $I = \{1,\ldots,r\}$. A particular partition $J^r = \{J_1,\ldots,J_r\}$ of the set J defines the following max-min type function:

$$\varphi(z) = \max_{i\in I}\min_{j\in J_i}(\langle x_j,z\rangle - y_j)\ \ (z \in \mathbf{R}^n). \tag{9}$$

We say that the sets A and B are *max-min separable* if there exist a finite number of hyperplanes, H, and a partition J^r of the set J such that for all $i \in I$ and $a \in A$ we have $(\langle x_j,a\rangle - y_j) < 0$ and for any $b \in B$ there exists at least one j such that $(\langle x_j,b\rangle - y_j) > 0$. It follows from the definition above that if the sets A and B are max-min separable then $\varphi(a) < 0$ for any $a \in A$ and $\varphi(b) > 0$ for any $b \in B$, where the function φ is defined by (9). Thus the sets A and B can be separated by a function represented as a max-min of linear functions. The problem of the max-min separability is reduced to the following optimization problem:

$$\min f(x,y) \quad \text{such that } (x,y) \in \mathbf{R}^{l\times n} \times \mathbf{R}^l, \tag{10}$$

where the objective function f is $f(x,y) = f_1(x,y) + f_2(x,y)$. Here, $f_1(x,y) = \frac{1}{m}\sum_{k=1}^m \max[0, \max_{i\in I}\min_{j\in J_i}(\langle x_j,a^k\rangle - y_j + 1)]$, $f_2(x,y) = \frac{1}{p}\sum_{s=1}^p \max[0, \min_{i\in I}\max_{j\in J_i}(-\langle x_j,b^s\rangle + y_j + 1)]$. The functions f_1 and f_2 are piece-wise linear, therefore the resulting function f is piecewise linear and consequently piecewise separable. In [2] it is shown that even for very simple cases these type of functions may not be regular and therefore the calculation of their subgradients is quite difficult. A

derivative-free algorithm for minimization of max-min type functions is proposed in [2]. This algorithm is the modification of the discrete gradient method. The results of the numerical experiments demonstrate that the algorithm is efficient for solving large scale problems up to 2000 variables.

4 Conclusion

This paper introduces some recent optimization methods developed in data mining by some modern areas of clustering and classification. There is a great potential of important OR applications, and of future research waiting.

References

1. E.J. Anderson and P. Nash, John Wiley and Sons Ltd, *Linear Programming in Infinite-Dimensional Spaces*, 1987.
2. Bagirov, A.M., and Ugon, J., *Piecewise partially separable functions and a derivative-free algorithm for large scale nonsmooth optimization*, Journal of Global Optimization 35 (2006) 163-195.
3. Bagirov, A.M., and Yearwood, J., *A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems*, EJOR 170, 2 (2006) 578-596.
4. Bock, H.H., *Automatische Klassifikation*, Vandenhoeck and Ruprecht, Göttingen (1974).
5. Cristianini, N., and Shawe-Taylor, J., *An introduction to Support Vector Machines and other Kernel-Based Learning Methods*, Cambridge University Press (2000).
6. Han, J., and Kamber, M., *Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems*, Jim Gray, Series Editor Morgan Kaufmann Publishers (2000).
7. Hastie, T., Tibshirani, R., and Freedman, J., *The Elements of Statistical Learning - Data Mining, Inference and Prediction*, Springer Series in Statistics, 2001.
8. Özöğür-Akyüz, S. and Weber, G.-W., *Learning with Infinitely Many Kernels via Semi-Infinite Programming*, in ISI Proceedings of 20th Mini-EURO Conference *Continuous Optimization and Knowledge-Based Technologies*, Neringa, Lithuania, May 20-23, 2007.
9. Sonnenburg, S., Raetsch, G., Schafer, C. and Schoelkopf, B. (2006), Large scale multiple kernel learning, J. Machine Learning Research 7, (2006) 1531-1565.
10. Volkovich, Z.V., Barzily, Z., Akteke-Öztürk, B., and Weber, G.-W., *Cluster stability using minimal spanning trees - a contribution to text and data mining*, submitted to TOP.
11. Weber, G.-W., Taylan, P., Özögür, S., and Akteke-Öztürk, B., Statistical learning and optimization methods in data mining, in: *Recent Advances in Statistics*, Turkish Statistical Institute Press, Ankara (2007) 181-195.