Discrete Applied Mathematics 165 (2014) 270–282

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Classifying negative and positive points by optimal box clustering

Paolo Serafini *

Università di Udine, Dipartimento di Matematica e Informatica, Via delle Scienze 206, 33100 Udine, Italy

ARTICLE INFO

Article history:
Received 23 November 2011
Received in revised form 23 October 2012
Accepted 2 May 2013
Available online 4 June 2013

Keywords:
Data analysis
Boxes
Patterns

ABSTRACT

In this paper, we address the problem of classifying positive and negative data with the technique known as box clustering. A box is homogeneous if it contains only positive (negative) points. Box clustering means finding a family of homogeneous boxes jointly containing all and only positive (negative) points. We first consider the problem of finding a family with the minimum number of boxes. Then we refine this problem into finding a family which not only consists of the minimum number of boxes but also has points that are covered as many times as possible by the boxes in the family. We call this problem the maximum redundancy problem. We model both problems as set covering problems with column generation. The pricing problem is a Maximum Box problem. Although this problem is NP-hard, there is available in the literature a combinatorial algorithm which performs well. Since the pricing has to be carried out also in the branch-and-bound search of the set covering problem, we also consider how the pricing has to be modified to take care of the branching constraints. The computational results show a good behavior of the set covering approach.

1. Introduction

A fundamental task in data analysis is the classification of data. Most often the information gathered from an individual (whatever it may be) is a set of real numbers, and the classification consists in deciding whether the data related to this individual belong to one class or to a complementary class. We may denote these two classes as 'positive' and 'negative'. There is a wide range of applications which can be framed into a model of this type.

Since the data related to an individual consist of an array of numbers, it is natural to associate to the individual a point in a vector space. If the data are meaningful for the investigation for which they are collected, it is expected that the points in the same class are more or less clustered and separated from the points of the other class, thus defining two regions of space. An initial known 'training' set of data is used to identify these regions of space. Then new data are classified on the basis of the defined regions.

One of the oldest classification methods is based on the idea that the positive and negative points can be separated by a hyperplane [7]. If this is the case, the hyperplane is first identified from the training set via mathematical programming techniques and then used for further classification of new data. The idea is simple and powerful. However, it is quite frequent that the data are not linearly separable. The use of nonlinear surfaces to separate points can be pursued (see again [7]). However, the choice of a suitable nonlinear function seems not only technically hard but also difficult to be justified. A piecewise linear separation has been proposed in [1] when the data are not linearly separable, but the method does not always produce satisfactory results.

* Tel.: +39 0432558442; fax: +39 0432558499.
E-mail address: paolo.serafini@uniud.it.

0166-218X/$ – see front matter
http://dx.doi.org/10.1016/j.dam.2013.05.003

Instead of defining a region of space according to the sign of a function (analytically or algorithmically defined), a different approach consists in defining a region as the union of simple and similar sets. In other words, a family of similar sets is defined such that each positive data element of the training set is covered by at least one member of the family while no negative data of the training set is covered by any set of the family. An analogous family is defined for the negative data.

For this approach to be useful, some requirements have to be considered. For instance, a trivial family consists of very small sets, each one covering exactly one point of the training set. Obviously, new data could be classified only if they are a replica of some data in the training set, and this situation is meaningless for the investigation. Intuitively, we would like to have large sets covering many points, so that new data which are close to some point in the training set can easily fit into the set. At the same time, the sets should be 'essential', in the sense that it is worth extending a set only if new points have to be covered. Having a family of large and essential sets is close to having a family with few sets. Hence we may consider the problem of covering the points with the minimum number of sets. However, it seems useful for the classification to have available some measure of 'trust' for the new data. To this aim not only we would like to cover the points with the minimum number of sets, but we would like also to have many points, i.e., many regions of space covered by more than one box. This kind of redundancy can constitute a measure of trust for the classification of new data.

The idea of covering the training set of points can in principle work with any family of subsets. However, subsets which would seem most suited for this analysis, like ellipsoids, present relevant computational problem. Perhaps the best family from the computational point of view is the family of boxes. Although boxes are far from being smooth sets (smoothness seems a natural requirement for identifying a region), they can be represented with a minimum amount of information, and this simplicity can lead to viable algorithms. Box clustering has been proposed and described in [6].

Moreover, there is a strong rationale behind the idea of using boxes. If we think of medical data, the result of an analysis is usually framed as a number belonging to some interval. The interval represents a set of values denoting a healthy status. Considering all data together, an individual is declared healthy if the point representing all measurements is within the high-dimensional box given by the Euclidean product of the single intervals. This is a simplified picture, since it assumes that the data are not correlated. In case they are, more than one box may be needed to cover the points. A more complex picture can be called for the data referring to sick individuals, since we do not expect these data to fit into one box.

In this paper, we want to find a covering of the positive (or the negative) points with the minimum number of boxes and also with maximum redundancy. The problem is modeled as a set covering problem with column generation, where each column is associated to a particular box. It turns out that the pricing problem consists in finding a single box of maximum weight. This last problem has been fully investigated in [4], and a viable combinatorial algorithm has been proposed. We base the set covering problem on this pricing algorithm. The problem of finding a minimum cardinality box cover in the bidimensional case has been investigated in [2], providing interesting complexity and algorithmic results.

In Section 2, we define the problem formally and provide some computational complexity results. In Section 3, we develop the set covering model with column generation. The corresponding pricing problem is described in Section 4, and the problem of introducing branching constraints in the pricing is dealt with in Section 5. In Section 6, we report some computational results applied to artificial data and to real data. Finally, in Section 7, we provide some conclusions.

2. Statement of the problem

There are given p *positive* points $X^i \in R^n$ $(i \in [p])$ and q *negative* points $Y^i \in R^n$, $(i \in [q])$. We use the notation $[h] := \{i \in Z : 1 \le i \le h\}$. We are interested in sets covering either only positive points or only negative points. In this paper, we consider only boxes as possible covering sets. Although models with different types of sets can be envisaged in general, we do not pursue this direction, which, very likely, entails considerable difficulties. A *box* $B(\ell, u)$ is the set

$$\left\{X \in R^n : \ell_k \le X_k \le u_k,\ k \in [n]\right\}.$$

A box is *positive* (*negative*) if it covers, i.e., it contains, only positive (negative) points. A box which is either positive or negative is called *homogeneous*. With abuse of notation, we denote by $|B|$ the number of points contained in the homogeneous box B.

A family of boxes $\left\{B_j\right\}_{j\in J}$ is *positive* (*negative*) if every positive (negative) point is contained in at least one positive (negative) box, i.e., a positive family $\left\{B_j\right\}_{j\in J}$ satisfies

$$X^i \in \bigcup_{j\in J} B_j, \quad i \in [p], \qquad Y^i \notin \bigcup_{j\in J} B_j, \quad i \in [q],$$

and a negative family $\left\{B_j\right\}_{j\in J'}$ satisfies

$$X^i \notin \bigcup_{j\in J'} B_j, \quad i \in [p], \qquad Y^i \in \bigcup_{j\in J'} B_j, \quad i \in [q].$$

Positive and negative families are called homogeneous.

For a given subset S of points, the *box closure* of S, denoted as $[S]$, is the smallest box containing all points in S, i.e.,

$$[S] := \left\{Z \in R^n : \min_{X\in S} X_k \le Z_k \le \max_{X\in S} X_k,\ k \in [n]\right\}.$$