A New Heuristic Algorithm for Consistent Biclustering

A. Mucherino[1], S. Cafieri[2]

[1] INRIA, Lille Nord Europe, France
[2] ENAC, Toulouse, France
antonio.mucherino@inria.fr
sonia.cafieri@enac.fr

1 Introduction

Data mining techniques have as aim to discover important information from large sets of data [4]. Generally, data related to a certain problem are collected and successively analyzed by such techniques. The set of data is basically formed by samples, which are represented by a sequence of features, that are considered to be relevant for the representation of the samples. The final problem that usually needs to be solved is to find a suitable partition of the samples of the set of data, where samples having similar properties are grouped together. In some applications, a training set, that is a subset of data for which a partition for the samples is already available, is known. When this is the case, the training set can be exploited for understanding how to perform the same classification on subsets of data which are not training sets. Classification techniques can be used for this purpose.

We consider a recently proposed technique for classification that is based on the concept of *consistent biclustering* [1]. Instead of considering only the samples in the sets of data, biclustering aims at finding simultaneous classifications of samples and of the features used for their representation. Moreover, if a training set is available, a biclustering can be constructed by exploiting this training set. The corresponding partition in biclusters is able to associate subgroups of samples to subgroups of features, so that the features causing the classification of the training set are revealed. This information can be therefore exploited for performing classification of samples which do not belong to the training set.

In order to perform correct classifications, it is very important that the considered biclusterings are *consistent* (see Section 2 for the definition of consistent biclustering). The problem of finding a consistent biclustering can be formulated as an optimization problem, which is NP-hard. We propose a new heuristic algorithm for the solution of this optimization problem. To this aim, we reformulate the problem as a bilevel optimization problem, where the inner problem is linear and hence solvable by standard software for linear optimization. We briefly discuss the concept of consistent biclustering and we present our heuristic algorithm in Section 2. Concluding remarks are given in Section 3.

2 A new heuristic algorithm

Let us suppose that a training set for a certain classification problem is available, that consists in a subset of data for which all samples already have a known classification in k disjoint classes: $B_S = \{S_1, S_2, \ldots, S_k\}$. Starting from this information, we can construct a classification for the features of the training set, by using the following procedure. For each class S_r in B_S, the average vector c_r^S corresponding to all the samples it contains can be computed. The vector c_r^S can be considered as the *center* of the class S_r. The i^{th} component c_{ir}^S of this vector represents the average expression of the i^{th} feature in its class: we can assign each feature of the set of data to the class where it is mostly expressed. In other words, if the i^{th} feature in the set is, in average, mostly expressed for the samples belonging to the r^{th} class of samples, then it is assigned to the r^{th} class

of features. Let $B_F = \{F_1, F_2, \ldots, F_k\}$ be the computed classification for the features in k classes. By combining the two classifications B_S and B_F, we can construct a biclustering of the training set: $B = \{(S_1, F_1), (S_2, F_2), \ldots, (S_k, F_k)\}$. Moreover, by using the same procedure, a new classification for the samples can be obtained from the classification B_F of the features: $\hat{B}_S = \{\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_k\}$. If B_S and $\hat{B}_S$ are the same, then, by definition, B is a consistent biclustering.

Unfortunately, real-life sets of data do not usually allow for consistent biclusterings. This is due to the fact that some features used for representing the samples actually do not represent the data very well. Such features need therefore to be removed from the set of data, while the total number of considered features is maximized in order to preserve the information in the training set. This problem can be formulated as a 0–1 linear fractional optimization problem, which is NP-hard. Some heuristic algorithms have been proposed for solving this optimization problem [1, 5].

The basic idea behind the heuristic algorithm that we propose is to reformulate the original 0–1 linear fractional optimization problem [1] into a bilevel program [3]. New real variables y_r, $\forall r \in \{1, 2, \ldots, k\}$ (where k is the number of biclusters), are defined and used together with the binary variables x_i, $\forall i \in \{1, 2, \ldots, m\}$ (where m is the number of features), of the original problem. The objective function of the bilevel program depends on both variables x_i and y_r, whereas only the x_i's are variables for the inner problem (the y_r's are considered as parameters). The objective function of the outer problem is able to provide a measure of how much the current biclustering is far from being consistent, and it is non-linear. The inner problem is a variant of the original optimization problem, and it is linear.

Our heuristic algorithm, which is based on the bilevel reformulation, takes inspiration from the Variable Neighborhood Search (VNS) [2], which is one of the most successful meta-heuristic searches for global optimization. At each iteration of the algorithm, values for the variables y_r are chosen by employing a random mechanism, while values for the variables x_i are obtained by solving exactly the inner optimization problem. In our experiments, the general VNS framework is implemented in AMPL, and the inner problem is solved by CPLEX at each iteration. This algorithm has been able to provide biclusterings of sets of data related to real-life applications, which have a higher quality if compared to the results obtained by other previously proposed heuristic algorithms [1, 5]. The reader who is interested in more details on the bilevel reformulation and on the proposed heuristic algorithm is referred to [3].

3 Conclusions

We presented a new heuristic algorithm for finding consistent biclusterings, which is based on a bilevel reformulation of the original optimization problem. We plan to employ this algorithm for solving classification problems arising in applied fields, such as agriculture, biology and chemistry.

References

[1] S. Busygin, O.A. Prokopyev, P.M. Pardalos. Feature Selection for Consistent Biclustering via Fractional 0-1 Programming. Journal of Combinatorial Optimization 10:7-21, 2005.

[2] P. Hansen, N. Mladenovic. Variable Neighborhood Search: Principles and Applications. European Journal of Operational Research 130(3):449–467, 2001.

[3] A. Mucherino, S. Cafieri. A New Heuristic for Feature Selection by Consistent Biclustering. arXiv e-print, arXiv:1003.3279v1, March 2010.

[4] A. Mucherino, P. Papajorgji, P.M. Pardalos. Data Mining in Agriculture. Springer, 2009. ISBN: 978-0-387-88614-5, 274 pages.

[5] A. Nahapatyan, S. Busygin, and P.M. Pardalos. *Mathematical Modelling of Biosystems*, chapter An Improved Heuristic for Consistent Biclustering Problems, pages 185–198. Applied Optimization 102, Springer, 2008. ISBN: 978-3-540-76783-1, 305 pages.