

# A Heuristic for Biclustering based on a Bilevel Reformulation

A. Mucherino<sup>1</sup>, L. Jourdan<sup>1,2</sup>, and E.-G. Talbi<sup>1,2</sup>

INRIA Lille Nord Europe, Villeneuve d'Ascq, France  
antonio.mucherino@inria.fr

Laboratoire d'Informatique Fondamentale de Lille, Bat M3 Cité Scientifique, Villeneuve d'Ascq, France  
{laetitia.jourdan,e1-ghazali.talbi}@lifl.fr

## 1 Introduction

Heuristic and meta-heuristic algorithms are widely used for solving various optimization problems, especially when there are no known exact algorithms for the considered problem, or when exact algorithms are too computationally expensive. Heuristics and meta-heuristics are often inspired by natural phenomena and animal behavior, or even by the problem to be solved [10]. They try to guide the search toward the optimal solutions of the considered problem, but, differently from exact algorithms, they cannot guarantee that the solution (or the set of solutions) of an optimization problem can be identified. Recently, therefore, the scientific community started to give attention to hybrid algorithms combining heuristics and exact algorithms, with the aim of developing algorithms having their best properties: the flexibility of heuristic algorithms, together with the reliability of exact algorithms [5, 9]. Naturally, an algorithm with these desirable properties together is quite difficult to develop, and hence a trade-off among such properties needs to be found.

We consider the data mining problem of finding a consistent biclustering of a given training set. Samples of a training set, together with the features used for representing the samples, can be organized on the rows and on the columns of a matrix  $A$ . A biclustering for  $A$  is a partition in disjoint submatrices of  $A$ , to which we refer as *biclusters*. Each bicluster is able to associate a subgroup of features to a subgroup of samples, hence revealing the features that cause the partition of the samples of the training set. This information can be exploited for attempting the reconstruction of similar partitions for sets of data which are not training sets: in other words, the found biclustering can be used for performing supervised classifications. The probabilities to perform correct classifications increases if the found biclustering is *consistent* (see next section). The problem of finding a consistent biclustering of a given training set can be formulated as an optimization problem, which is NP-hard [6]. We provide some more details on this problem in Section 2.

We present a hybrid heuristic algorithm for solving this optimization problem. The algorithm is based on a reformulation of the problem as a *bilevel program*, i.e. as an optimization problem (the outer problem) in which one of its constraints is the solution of another optimization problem (the inner problem). The presented algorithm is composed by a heuristic framework for the solution of the outer problem in which, at each iteration, the inner problem is solved by an exact algorithm. Section 3 is devoted to a brief presentation of this hybrid heuristic algorithm. Conclusions are given in Section 4.

## 2 Classification by Consistent Biclustering

Given a set of data  $A$ , biclustering aims at finding simultaneous partitions of the samples and of the features that are used for their representation. If  $A$  is a training set, a biclustering for  $A$  can be obtained by employing the supervised technique described in [2]. The basic idea behind such a technique is the following. For each class of the training set, one bicluster is created and all the samples belonging to one class are associated to it. Then, each feature is associated to the bicluster containing the samples where, in average, the feature is mostly expressed. We point out that this technique can also be inverted: a possible partition for the samples of  $A$  can be computed by using a known partition of its features. By definition, a biclustering is consistent when both partitions of samples and of features can be correctly reconstructed by this technique. For a discussion on consistent biclusterings, the reader is referred to [2, 7].

Real-life sets of data do not usually allow for consistent biclusterings, because some of the used features may actually not represent well the data. Therefore, we need to remove such features 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 [6]. Some heuristic algorithms have been proposed for the solution of this problem.

### 3 A Hybrid Heuristic Algorithm

The 0–1 linear fractional optimization problem which needs to be solved for finding consistent biclusterings of training sets has nonlinear constraints [2], and only heuristic algorithms have been proposed so far for its solution. In the bilevel formulation (the reader is referred to [7] for a formal definition of the bilevel problem that we need to omit here for lack of space), the outer problem is still nonlinear, whereas the inner problem is linear and therefore solvable by standard algorithms for linear optimization. This inspired the development of a hybrid algorithm for the solution of the problem.

The optimization of the outer problem is based on a heuristic framework which borrows ideas from the meta-heuristic Variable Neighborhood Search (VNS) [3]. The basic idea is to focus the search for better-quality solutions in small neighbors of the current solution, and to enlarge such neighbors only when no better solutions are found. At each iteration of the heuristic framework, the inner problem is solved exactly (it changes at each iteration because it depends on some decision variables of the outer problem). This allows to randomly select only one part of the decision variables of the problem, while the other part is selected by solving exactly the inner problem. More details regarding the hybrid algorithm can be found in [7].

The presented algorithm has been implemented in AMPL [1], where the inner problem has been solved at each iteration by CPLEX [4]. It has already been used in [7] for analyzing gene expression data. The hybrid algorithm has been able to identify the genes (features) involved in certain diseases, so that accurate predictions of such diseases can be performed by analyzing a subset of gene expressions. Moreover, in [8], the hybrid algorithm has been used for analyzing wine fermentations. Compounds of wine (features) have been measured regularly for monitoring normal and problematic fermentations. The found biclusterings revealed the major compounds causing problematic fermentations.

### 4 Conclusions

We discussed a hybrid heuristic algorithm for finding consistent biclusterings of training sets. This is a very important problem in data mining, because consistent biclusterings can be exploited for solving classification problems. The hybrid algorithm is based on a bilevel reformulation of the original problem, in which the inner problem is linear. The basic idea is to implement a heuristic strategy for finding the optimal values of only one part of the decision variables, while the values for the other variables are found by solving the inner problem exactly. We believe that the same idea could be exploited for solving other NP-hard optimization problems.

### References

1. AMPL, <http://www.ampl.com/>
2. 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.
3. P. Hansen, N. Mladenovic, *Variable Neighborhood Search: Principles and Applications*, European Journal of Operational Research **130** (3), 449–467, 2001.
4. ILOG, CPLEX, <http://www.ilog.com/products/cplex/>
5. L. Jourdan, M. Basseur, and E-G. Talbi, *Hybridizing Exact Methods and Metaheuristics: A Taxonomy*, European Journal of Operational Research **199** (3), 620–629, 2009.
6. O.E. Kundakcioglu, P.M. Pardalos, *The Complexity of Feature Selection for Consistent Biclustering*, In: Clustering Challenges in Biological Networks, S. Butenko, P.M. Pardalos, W.A. Chaovalitwongse (Eds.), World Scientific Publishing, 2009.
7. A. Mucherino, S. Cafieri, *A New Heuristic for Feature Selection by Consistent Biclustering*, arXiv e-print, arXiv:1003.3279v1, March 2010.
8. A. Mucherino, A. Urtubia, *Consistent Biclustering and Applications to Agriculture*, IbaI Conference Proceedings, Proceedings of the Industrial Conference on Data Mining (ICDM10), Workshop on Data Mining and Agriculture (DMA10), Berlin, Germany, 105-113, 2010.
9. E-G. Talbi, *A Taxonomy of Hybrid Meta-Heuristics*, Journal of Heuristics **8**(2), 541–564, 2002.
10. E-G. Talbi, *Metaheuristics. From Design to Implementation*, John Wiley & Sons, 2009.