Consistent Biclustering and Applications to Agriculture

A. Mucherino[1] and A. Urtubia[2]

[1] INRIA Lille Nord Europe, Villeneuve d'Ascq, France,
antonio.mucherino@inria.fr
[2] Departamento de Ingenieria Quimica y Ambiental,
Universidad Tecnica Federico Santa Maria, Valparaiso, Chile,
alejandra.urtubia@usm.cl

Abstract. Consistent biclusterings of training sets can be exploited for solving classification problems in data mining. This technique has been mainly applied so far to solve classification problems related to gene expression data. However, it can be successfully applied to problems arising in other domains, and it is also able to provide information on the features causing the classification of the training set. We provide a quick overview of this technique, and we present a study on a particular problem arising in the agricultural field. We consider the problem of predicting problematic fermentations of wine at early stages of the vinification. The presented computational experiments show that the considered technique is able to provide some clues on the possible features causing the problematic fermentations.

1 Introduction

Techniques for mining data are more and more of interest because of the growing amount of information available from different resources which needs to be analyzed. In many real-life applications, the quantity of information is huge, and the problem is to extract from the *ocean* of available data only important information. Discovering such information has mainly two practical consequences. First, relationships among the data can be found, so that a prediction of data which are otherwise difficult to obtain can be attempted by exploiting the acquired knowledge. Secondly, before storing data into a database, only the pertinent and relevant information can be identified by data mining techniques.

Nowadays, data mining techniques are widely used for analyzing data from real-life applications [10]. Given a set of data, such techniques have as final aim to find a partition of the set of data, where similar data are grouped together. This partition of the data allows for discovering important relationships. In some applications, a subset of data which is already divided in subgroups is available, and such subset can therefore be exploited for understanding how to perform similar classifications on other subsets related to the same problem. This subset is referred to as *training set*, and classification techniques use it in order to *learn* how to solve classification problems. When no training sets are available,

clustering techniques are employed, which are able to work with sets of data for which no other information are known. It is important to note that there are techniques for clustering, such as techniques for hierarchical clustering, that do not provide only a partition of the data, but also complementary information. A recent survey on data mining techniques can be found in [11].

The focus of this paper is on classification techniques, and, in particular, on a novel classification technique which is based on the concept of *consistent biclustering* [3]. It is important to note that, even though the considered technique constructs biclusterings of a set of data, it is actually a technique for classification (that exploits the information from a training set). As a consequence, typical expressions employed when dealing with classification techniques, such as "classes of samples", and with clustering techniques, such as "partition in (bi)clusters", will be both used in the following discussion.

A 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. Instead of considering samples only, biclustering aims at finding simultaneous classifications of samples and of their features. Moreover, if a training set is known, 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 then be 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 found biclustering is consistent (see Section 2 for the definition of consistent biclustering). However, 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 [7]. Some heuristic algorithm have been proposed for solving this optimization problem, and, in the experiments presented in this paper, we will consider the heuristic algorithm recently proposed in [9].

This technique has been used so far for analyzing gene expression data [3, 12], where samples may represent diseases, human tissues, etc., but also for studying brain dynamics in patients affected by epilepsy [2]. In this paper, we use the technique for analyzing different wine fermentations by using data experimentally measured during the first 150 hours of the fermentation process. Our main aim is to predict problematic fermentations in time for an enologist to interfere with the process and ensure that the fermentation could end regularly and smoothly. Compounds are regularly measured from different wine fermentations of the *Cabernet sauvignon*. A training set of wine fermentations is defined and studied by the considered technique. As shown in Section 3, some features causing problematic fermentations are identified by constructing biclusterings of the training set.

The rest of the paper is organized as follows. In Section 2 we will describe the considered technique for consistent biclustering, and we will briefly present the heuristic algorithm that we use for the solution of the corresponding 0–1 linear fractional optimization problem. In Section 3 we will present our experiments for the analysis of a set of data related to different fermentations of wine, and we will discuss the obtained results. Conclusions will be given in Section 4.

2 Classification by consistent biclustering

Let A be the training set for a certain classification problem. This set of data can be seen as a matrix A, whose columns a^j represent samples and whose rows a_i represent features. Hence, the generic element a_{ij} of the matrix corresponds to the i^{th} feature of the j^{th} sample. A classification is associated to the columns of A. If n is the number of samples, and m is the number of features, then A is a $m \times n$ matrix. The basic idea behind the considered classification technique is to find a biclustering for A which is consistent [3, 4]. The consistent biclustering is able to associate relevant features to each class of the classification problem, and this information can be used for classifying samples which do not belong to the training set.

The first step is therefore to construct a biclustering from the available training set A, i.e. a partition in disjoint sub-matrices of A. Since A is a training set, a classification for its samples in k classes is known:

$$B_S = \{S_1, S_2, \ldots, S_k\}.$$

Starting from this classification, a classification for the features can be constructed as follows. 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 component c_{ir}^S of this vector represents the average expression of the feature a_i in its class: we can assign each feature of A to the class where it is mostly expressed (for a formal definition of this procedure, the reader is referred to [3]). In other words, we can create a classification for the features by assigning each feature a_i to the class $F_{\hat{r}}$ if and only if $S_{\hat{r}}$ is the class of samples where this feature is mostly expressed, i.e.:

$$c_{i\hat{r}}^S > c_{ir}^S \qquad \forall r \neq \hat{r}.$$

Let

$$B_F = \{F_1, F_2, \ldots, F_k\}$$

be the computed classification for the features in k classes. Starting from this classification, we can obtain another classification for the samples:

$$\hat{B}_S = \{\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_k\}$$

by applying the same procedure, where the average expressions of the samples in the known classes of features are considered. In general, the two classifications

B_S and $\hat{B}_S$ are different from each other. If they coincide, then the biclustering

$$B = \{(S_1, F_1), (S_2, F_2), \ldots, (S_k, F_k)\}$$

is, by definition, a *consistent biclustering* [3].

As already remarked, training sets related to real-life applications do not usually allow for consistent biclusterings if all their features are considered. Therefore, a subset of relevant features needs to be found, and this *feature selection* problem can be formulated as the following optimization problem [3]:

$$\max_x \left(f(x) = \sum_{i=1}^{m} x_i \right) \tag{1}$$

subject, $\forall \hat{r}, \xi \in \{1, 2, \ldots, k\}, \hat{r} \neq \xi, j \in S_{\hat{r}}$, to:

$$\frac{\sum_{i=1}^{m} a_{ij} f_{i\hat{r}} x_i}{\sum_{i=1}^{m} f_{i\hat{r}} x_i} > \frac{\sum_{i=1}^{m} a_{ij} f_{i\xi} x_i}{\sum_{i=1}^{m} f_{i\xi} x_i}, \tag{2}$$

where:

- x_i is a binary variable: it is 1 if the feature a_i is selected, and it is 0 otherwise;
- f_{ir} is a binary parameter: it is 1 if the i^{th} feature belongs to the class F_r, and it is 0 otherwise.

The objective function (1) of this optimization problem is simply the counter of features that are selected, which is maximized in order to preserve the information in the set of data. Note that the two fractions in the generic constraint (2) are used for computing the average expressions of samples in classes of features. Such a constraint ensures that the $\hat{r}$-th sample is the mostly expressed if it belongs to the bicluster $(S_{\hat{r}}, F_{\hat{r}})$. As a consequence, if all the constraints are satisfied, the corresponding biclustering is consistent. This is a 0–1 linear fractional optimization problem, which is NP-hard [7].

It is important to note that two other similar optimization problems have been proposed for overcoming some issues related to noisy data and experimental errors that may affect the data. In practice, these two optimization problems have the same formulation of problem (1)-(2), but an additive or multiplicative parameter is added to the second fraction of the constraints (2). In this way, the margin between biclusters is enlarged, and variations in the data due to noise or experimental errors are less likely able to cause misclassifications. We refer to α-*consistent biclustering* when an additive parameter $\alpha > 0$ is used, and to β-*consistent biclustering* when a multiplicative parameter $\beta > 1$ is used. The reader is referred to [12, 13] for additional details on these optimization problems. In the following discussion, we will consider only the optimization problem (1)-(2), because similar observations can be made for the other two. The experiments

that we will present in the next section, however, will be related to all three optimization problems.

Heuristic algorithms [3, 9, 12] have been proposed for the solution of the optimization problem (1)-(2). In this paper, we will consider the heuristic algorithm recently proposed in [9], that has been proved to be more efficient than the previous ones on some sets of data related to real-life applications. We will not go in the details of this heuristic algorithm, but we will rather refer the interested reader to [9] for a wide discussion on the used algorithm. We only mention that the considered heuristic algorithm is based on a bilevel reformulation of the optimization problem (1)-(2), in which the inner problem is linear. The basic framework of the heuristic algorithm is taken from the meta-heuristic Variable Neighborhood Search (VNS) [5, 8], which is one of the best meta-heuristic algorithms for global optimization. At each iteration of the algorithm, however, the linear inner problem is solved exactly by CPLEX [6]. We implemented this heuristic algorithm in AMPL [1], and we used it for finding consistent, α-consistent or β-consistent biclusterings of the set of data related to wine fermentations.

3 Prediction of problematic wine fermentations

Problems occurring during the fermentation process of wine can impact the productivity of wine-related industries and also the quality of wine [11]. The fermentation process of wine can be too slow or it can even become stagnant. Predicting how good the fermentation process is going to be may help enologists who can then take suitable steps to make corrections when necessary and to ensure that the fermentation process concludes smoothly and successfully. In order to monitor the wine fermentation process, various compounds can be measured and the data obtained during the fermentation process can be analyzed [14–16]. Data mining techniques can help extracting this information from large sets of data. Such information can then be used to predict the quality of fermentation processes.

We present some analysis performed on a set of data obtained from a winery in Chile's Maipo Valley, which is the result of 24 measurements of industrial vinifications of *Cabernet sauvignon* [14]. The data are related to the harvest of 2002. Between 30 and 35 samples are taken per fermentation depending on the duration of the analyzed vinification. The level of 29 compounds are analyzed. Among them, sugars are analyzed, such as glucose and fructose, and also organic acids, such as the lactic and citric acids, and nitrogen sources, such as alanine, arginine, leucine, etc., and alcohols (all the details regarding the experimental analysis can be found in [14]). The whole set of data consists of approximately 22000 data points. In this work, the used compounds are actually 30, because we added one for representing the *total sugar*, which consists of the sum of the glucose and fructose levels. In general, the added variable is considered to contain redundant information, but, since the used technique is able to select the most relevant features, we left this task to it.

Among the 24 fermentations, 9 fermentations ended normally, whereas the other 15 had problems. Fermentations that finished with more than 2 g/L of

α	0.00	0.10	0.20	0.40	0.50	0.70	1.00
$f(x)$	192	192	192	190	190	170	149
β	1.00	1.01	1.02	1.04	1.05	1.07	1.10
$f(x)$	192	191	186	186	180	177	165

Table 1. Computational experiments on the set of wine fermentations. The features are selected by finding consistent, α-consistent or β-consistent biclusterings.

sugar or lasted more than 13 days are arbitrarily defined as problematic. In particular, 10 problematic fermentations were too slow (they would be able to end correctly but in a longer time), whereas 5 problematic fermentations got stuck (there is no more yeast activity, and hence sugar is not anymore transformed in alcohol). In the following analysis, each fermentation is represented by a sequence of measurements of the 30 compounds at different stages of the fermentation process. No data after the first 150 hours of fermentation are considered, in order to verify the capability to predict the problematic fermentations at early stages. Measurements in the sequences are sorted from the first to the last, and, for each sample, measurements occurring approximately at the same time take the same position in the sequence. For one (slow) fermentation, only 5 measurements are available after 150 hours, and therefore we removed it from our set of data. In general, at least 8 measurements are available for all the other fermentations. Finally, the considered set of data contains 23 fermentations. These 23 fermentations are classified in 3 different classes: the first class contains normal fermentations (9 in total), the second class contains slow fermentations (9 in total), and the third class contains fermentations got stuck (5 in total). Each sample of the set is formed by a sequence of 8 measurements consisting in the levels of 30 compounds. In total, each sample is represented by 240 features.

Table 1 shows the consistent, α-consistent and β-consistent biclusterings that the considered heuristic algorithm has been able to find. As expected, the number of selected features decreases when the value given to the parameter α or β is larger, because the margin among the classes is enlarged for avoiding issues related to experimental errors and noise. When α-consistent and β-consistent biclusterings are searched, we are able to select the features that are actually relevant for the three classes of our training set. However, it is important to note that the optimization problems related to α-consistent or β-consistent biclusterings become more and more difficult to solve when α or β is increased. Hence, a trade-off needs to be found.

In this paper, we provide details for only one of the found biclusterings, and in particular for the one where the smallest number of features (149) has been selected, because β was set to 1.10. It is reasonable to believe that the most relevant features are the ones selected in this biclustering. By analyzing the biclustering, we can note that not all the features related to the sugar levels are essential for the classification, because the features measured at the very early

stages of the fermentation were all discarded. However, after approximately 100 hours of fermentation, such features become important: they are all selected, and assigned to one of the two classes containing problematic fermentations. This result agrees with the fact that the problematic fermentations are the ones in which the sugar levels are higher. Unfortunately, all the selected features which are related to sugar levels seem to be randomly assigned to slow and stuck fermentations. As a consequence, these features are able to provide information on problematic fermentations, but they cannot be used for discriminating between slow and stuck fermentations.

Among the organic acids, the features related to lactic, malic, succinic, and tartaric acids are all preserved during the feature selection. Moreover, all the features related to each of these organic acids are assigned to only one bicluster: this shows that they can play a very important role for the classification of the fermentations. For example, the lactic acid is strongly related to the bicluster of stuck fermentations, and thus all fermentations with high levels of lactic acid are much likely going to get stuck. Moreover, the information on the levels of lactic acid seem to be very relevant starting from the first hours of the fermentation process. Regarding the malic acid, instead, all the features representing the measurements of this acid are strongly related to the bicluster containing the normal fermentations. As a consequence, fermentations in which the levels of this acid are higher should all end normally. Similarly, the features related to the succinic acid are all selected as well and they are all assigned to the bicluster of slow fermentations. Finally, the features related to the tartaric acid are all selected and they are all assigned to the bicluster of stuck fermentations.

Besides the information on the features that are always selected and that are able to represent well the bicluster to which they belong, it is also important to see which features are never selected. The features related to the amino acid arginine, for example, are always discarded, and therefore they can be completely removed from the set of data because they are not relevant for the classification of the fermentations. There are also other features related to the different measurements of the same compound that are always discarded. Examples are the proline, the glutamic acid, the glutamine, and the treonine. Other features related to the same compound seem to be selected and discarded with irregular patterns (differently from the case of the sugar levels where they are discarded at the very early stages only). We will not reveal in this paper the compounds for which we obtain these apparently irregular patterns: we plan to perform further deeper analyses on the temporal properties of the features related the same compound, and to provide more details on our experimental results in future publications.

4 Conclusions

We presented an analysis of a set of data related to normal and problematic wine fermentations with the aim of discovering important information that can be used for prediction purposes. We constructed various biclusterings of the set of

data by selecting the features, used for representing the fermentation processes, that are relevant for the classification of the fermentations. To this aim, we solved different feature selection problems by applying a recently proposed heuristic algorithm that is based on a bilevel reformulation of the original feature selection problem. Our preliminary experiments showed which features (levels of sugar, lactic acid, malic acid, succinic acid, and tartaric acid) are very important for the classification, and which features (levels of arginine, proline, glutamic acid, glutamine, and treonine) are not.

Future works will be aimed at the analysis of the entire set of data. Indeed, some compounds that apparently cannot give a good contribution to the classification problem might be useful when measured after the first 150 hours of fermentation. Obviously, this information cannot be exploited for improving the fermentation process at early stages, but it could be able to provide some insights on the vinification. In order to consider the entire set of data, however, we need to work on the considered heuristic algorithm and improve its performances. Indeed, if all the 22000 available measurements need to be considered, the corresponding feature selection problem is huge in size. We will also work in order to perform automatic predictions of the wine fermentations by using the found consistent biclusterings. We hope we will be able to show the results of these experiments in future publications.

Acknowledgments

The authors wish to thank the anonymous referees, whose comments helped us improving this paper.

References

1. AMPL, `http://www.ampl.com/`
2. S. Busygin, N. Boyko, P.M. Pardalos, M. Bewernitz and G. Ghacibeh, *Biclustering EEG Data from Epileptic Patients Treated with Vagus Nerve Stimulation*, AIP Conference Proceedings **953**, Data Mining, System Analysis and Optimization in Biomedicine, 162–173, 2007.
3. 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.
4. S. Busygin, O.A. Prokopyev, P.M. Pardalos, *Biclustering in Data Mining*, Computers & Operations Research **35**, 2964–2987, 2008.
5. P. Hansen, N. Mladenovic, *Variable Neighborhood Search: Principles and Applications*, European Journal of Operational Research **130** (3), 449–467, 2001.
6. ILOG, CPLEX, `http://www.ilog.com/products/cplex/`
7. 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.
8. M. Mladenovic, P. Hansen, *Variable Neighborhood Search*, Computers and Operations Research **24**, 1097–1100, 1997.

9. A. Mucherino, S. Cafieri, *A New Heuristic for Feature Selection by Consistent Biclustering*, arXiv e-print, arXiv:1003.3279v1, March 2010.
10. A. Mucherino, P. Papajorgji, P.M. Pardalos, *Data Mining in Agriculture*, Springer, 2009.
11. A. Mucherino, P. Papajorgji, P.M. Pardalos, *A Survey of Data Mining Techniques Applied to Agriculture*, Operational Research: An International Journal **9**(2), 121–140, 2009.
12. A. Nahapatyan, S. Busygin, and P.M. Pardalos, *An Improved Heuristic for Consistent Biclustering Problems*, In: Mathematical Modelling of Biosystems, R.P. Mondaini and P.M. Pardalos (Eds.), Applied Optimization **102**, Springer, 185–198, 2008.
13. P.M. Pardalos, O.E. Kundakcioglu, *Classification via Mathematical Programming*, Journal of Computational and Applied Mathematics **8** (1), 23–35, 2009.
14. A. Urtubia, J.R. Perez-Correa, M. Meurens, E. Agosin, *Monitoring Large Scale Wine Fermentations with Infrared Spectroscopy*, Talanta **64** (3), 778-784, 2004.
15. A. Urtubia, J.R. Perez-Correa, A. Soto, P. Pszczolkowski, *Using Data Mining Techniques to Predict Industrial Wine Problem Fermentations*, Food Control **18**, 1512–1517, 2007.
16. A. Urtubia, J.R. Perez-Correa, *Study of Principal Components on Classifications of Problematic Wine Fermentations*, Lecture Notes in Computer Science **5633**, 38–43, 2009.