FEATURE SELECTION FOR DATASETS OF WINE FERMENTATIONS

Antonio Mucherino[(a)], Alejandra Urtubia[(b)]

[(a)] CERFACS, Toulouse, France
[(b)] Universidad Tecnica Federico Santa Maria, Valparaiso, Chile

[(a)]mucherino@cerfacs.fr, [(b)]alejandra.urtubia@usm.cl

ABSTRACT
The fermentation is the most important process in the production of wine. Problematic fermentations can cause losses to wine makers, because such fermentations could be too slow to provide the final product, or they may even become stagnant. An efficient prediction of problematic fermentations at the first stages of the process is therefore of great interest. The aim of this paper is two-fold. We apply a supervised biclustering technique to a dataset of wine fermentations with the aim of selecting and discovering the features that are responsible for the problematic fermentations. We also exploit the selected features for predicting the quality of new fermentations, and we propose a new strategy for validating the obtained classifications.

Keywords: wine fermentations, feature selection, classification, combinatorial optimization

1. INTRODUCTION

Wine is widely produced around the world. Industrial production of wine is an important business in many countries. For this reason, the study of the fermentation process, which is able to transform grape juice into the alcoholic beverage, is of increasing interest in the field of agriculture. Problematic fermentations, indeed, may cause losses to industries. If a fermentation process is slower than usual, for example, the final product is produced in a longer time. Moreover, in the worst case, when the fermentation process gets stuck, a part of the production could be completely spoiled.

Data mining is a field of operations research that analyzes large databases with the aim of acquiring novel knowledge (Mucherino, Papajorgji, Pardalos, 2009). In recent years, data mining techniques have specifically been applied to agricultural problems in order to find important information about the problem under study. In the case of wine fermentations, a database of compound measurements, taken at different times during the fermentation process, can be exploited for extracting information that can help the prediction of problematic fermentations. This prediction could allow wine makers to interfere with the problematic fermentation processes in order to guaranteeing a good fermentation.

We present in this paper some analysis regarding wine fermentations. We consider a dataset obtained from a winery in Chile's Maipo Valley, which is the result of 24 measurements of industrial vinifications of Cabernet sauvignon (Urtubia, Perez-Correa, Meurens, 2004). By using a supervised biclustering technique, we select the compounds that are responsible for the problematic fermentations, and we also attempt a prediction of other unknown fermentations. In the case in which the fermentations are only divided in two classes (normal and problematic fermentations), we show that correct predictions can be performed by using the new classification strategy we present.

The rest of the paper is organized as follows. In Section 2, we discuss the problem of predicting wine fermentations, and we give more details about the considered dataset containing fermentations. In Section 3, we briefly present the data mining technique that we employ for performing wine predictions. In particular, we consider a supervised biclustering technique that is able to solve two problems at the same time (Busygin, Prokopyev, Pardalos, 2005). We can select the compounds measured during the fermentations that actually allow discriminating between normal and problematic fermentations, and we can also exploit the acquired knowledge for performing predictions on new fermentations. The basic strategy, as well as an improved strategy that we propose for performing the predictions, are both described in Section 4. Finally, Section 5 presents some experiments, and Section 6 concludes the paper.

2. PREDICTING WINE FERMENTATIONS

The problem of predicting wine fermentations is very important for wine makers. Some problems occurring during the fermentation process of wine can indeed impact the quality and the productivity of the final product (Mucherino, Papajorgji, Pardalos, 2009). 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. Classic problematic fermentations are slow fermentations, and fermentations that can get stuck at a certain point of the process.

We consider in this paper a dataset of wine fermentations that have been obtained from 24 measurements of industrial vinifications of Cabernet sauvignon (Urtubia, Perez-Correa, Meurens, 2004), in a winery in Chile's Maipo Valley. The data are related to the harvest of 2002. The levels 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. 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.

Measurements are sorted in a sequence from the first to the last. For each fermentation, measurements occurring approximately at the same time take the same position in this sequence. To this purpose, 15 temporal windows comprising 10 hours are defined, to which a subset of measurements are associated. When data in the temporal window were not available for some fermentation, they have been obtained by exploiting measurements close in time and by employing linear regression techniques. Finally, the considered set of data contains 24 fermentations described by $15 \times 30 = 450$ features: the first class contains *normal* fermentations (9 in total), whereas the second class contains *problematic* fermentations (15 in total). Preliminary studies on this dataset have been presented in (Mucherino, Urtubia, 2010).

In order to verify the quality of the performed predictions, this dataset has been divided into training and testing set. The training set contains 16 randomly chosen fermentations: 6 normal fermentations and 10 problematic fermentations. The testing set is smaller in size and it contains 8 fermentations in total: 3 normal and 5 problematic fermentations.

In this work, we consider the training set for selecting the features that allow for performing correct classifications of the fermentations. To this aim, we search *consistent biclusterings* of the training set, which are able to associate subgroups of features to subgroups of samples of the dataset (each sample represents one fermentation). In order to obtain a consistent biclustering of the training set, some features are removed from the set. This is done by solving a combinatorial optimization problem. Details about this procedure are given in Section 3.

Once a consistent biclustering is found from a training set, the corresponding relationship between samples and features can be exploited for classification purposes. Given a testing set related to the same problem, the classification of its samples, by definition, is supposed to be not known. However, the classification of its features is known, because it is exactly the same of the training set, and this information can therefore be exploited for reconstructing the classification of the samples of the testing set. More details are given in Section 4.

3. FEATURE SELECTION BY CONSISTENT BICLUSTERING

Let $A = \{a_{ij}\}$ be a matrix representing a dataset. The matrix A contains n samples (column by column) and m features (row by row). A *bicluster* is a submatrix of A, that is able to group together a subset of samples (a class S_r) and a subset of features (a class F_r) of the dataset. A biclustering

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

is a partition of A in disjoint biclusters that covers A (Busygin, Prokopyev, Pardalos, 2005).

Biclusterings of datasets are usually searched by unsupervised techniques, where it is supposed that no information about the data is available. The interested reader can read the survey by Madeira and Oliveira (2004) for a wider discussion on the topic. In our approach, instead, it is supposed that the set of data A is a training set, i.e. A is a set for which the classification of its samples is already known. The corresponding biclustering is therefore computed by employing a supervised technique, which is able to select the important features of the set. This information is then exploited for classifying samples having no known classification.

Let us suppose that A is a training set. Therefore, the classification of its samples in k classes is known. From this classification, it is possible to construct a classification of its features in k classes. The basic idea is to assign each feature to the class where it is mostly expressed (in other words, where it has higher value), in average, in the corresponding class of samples. Note that the same procedure can be inverted and it can be used for finding a classification of the samples of A from a known classification of its features. The reader can refer to Mucherino and Cafieri (2010) for more details about this supervised technique.

By combining the two classifications, the one for the samples of A, and the other one for the features of A, a biclustering can be defined for the matrix A. As already remarked, the supervised procedure mentioned above can construct classifications for the samples from classifications for the features, and vice versa. If the biclustering remains unchanged when the supervised procedure is applied, then it is said to be *consistent*. In other words, the biclustering is consistent if the classification of the samples (the features) suffices for

correctly reconstructing the biclustering (Busygin, Prokopyev, Pardalos, 2005).

In real life applications, however, training sets usually do not allow for consistent biclusterings. This is due to the fact that some features used for representing the samples could actually not be adequate for their representation. This is a common problem in data mining, because, at the time features are chosen for representing the samples, the knowledge about the problem is very limited. For this reason, in order to allow for consistent biclusterings, a subset of features must be removed from the training set. During this process, the number of rejected features must be as small as possible, in order to preserve the information contained in the dataset. This is done by solving a combinatorial optimization problem, where selected features are maximized while constraints ensure that the corresponding biclustering is consistent. The reader can find the formal definition of this optimization problem in (Busygin, Prokopyev, Pardalos, 2005).

In order to overcome issues related to noisy data and experimental errors, the concepts of α-consistent and β-consistent biclustering have been introduced in (Nahapatyan, Busygin, Pardalos, 2008). In order to find α-consistent or β-consistent biclusterings, the formulation of the optimization problem is slightly modified, and it allows for selecting the features that are relevant for the problem and, at the same time, the ones that should not be sensitive to noise and experimental errors.

Solving these three optimization problems (one for finding consistent biclusterings of training sets, one for α-consistent biclusterings and another for β-consistent biclusterings) is NP-hard (Kundakcioglu and Pardalos, 2009). In this work, we find approximations of the solutions to these problems by employing the VNS-based heuristic proposed in (Mucherino, Cafieri, 2010).

4. PERFORMING CLASSIFICATIONS WITH THE SELECTED FEATURES

We briefly described in the previous section a technique for selecting the features of *A* by finding a consistent biclustering of the training set. By exploiting this acquired knowledge, supervised classifications of the samples that do not belong to *A* can then be performed.

Let *B* be a dataset which is not a training set and that it is related to the same classification problem as the set *A*. No information regarding the classification of the samples in *B* is therefore available, but *B* contains the same features of *A* and a classification of these features is known because a biclustering for *A* is available. By using the supervised procedure, then, a classification for the samples of *B* can be found from the known classification of its features. Since the biclustering of *A* is consistent, the procedure is able to find the correct classification for the samples in *A*, and therefore it should be able to do most likely the same for the samples in *B* (Busygin, Prokopyev, Pardalos, 2005).

The main issue concerning this technique is related to the fact that the biclustering associated to *B* is most likely not consistent. When this is the case, there are probably samples in *B* that are misclassified. In practice, samples in *B* may contain some additional information that *A* did not contain, causing in this way the loss of consistency.

We propose therefore a new strategy that is able to manage the case in which the biclustering associated to *B* is not consistent. In order to verify which samples in *B* might be misclassified, we build and check the consistency of the biclusterings associated to a set of matrices A_1, A_2, …, A_b, where b is the number of samples in *B* and each A_j is defined by adding the j^{th} sample to *A*. If the biclustering associated to A_j is consistent, then the j^{th} sample is probably well classified by the supervised procedure.

If the biclustering associated to A_j is instead not consistent, the obtained prediction could be not correct. In general, we cannot immediately declare a certain prediction wrong if the biclustering associated to A_j is not consistent. However, we can try to verify the quality of these classifications.

Let us suppose that the prediction for the j^{th} sample suggests that it belongs to class 1 and that A_j is not consistent. We can therefore try to make this biclustering consistent by removing some other features that were not unselected during the solution of the optimization problem. To this aim, a similar optimization problem can be solved, where the matrix A_j is considered and previously unselected features are directly discarded. It is important to note that, in the definition of the problem, a classification for the j^{th} sample must be provided.

One possibility is to give to the j^{th} sample the same classification that the procedure suggested. In this case, however, since the given classification for the j^{th} sample is exactly what the prediction gives, the biclustering may become easily consistent. This does not ensure that the j^{th} sample actually belongs to class 1.

The second possibility is to give to this sample the opposite classification (say class 2). We can then solve the optimization problem for removing some further features. Solutions to this problem can give clues about the actual classification for the sample.

If many features are removed from A_j in order to make its corresponding biclustering consistent, then the j^{th} sample probably belongs to class 1. The fact that many features are discarded indicates that assigning the sample to the other class goes against the original information in *A*. When instead few more features are

removed from A_j when solving this optimization problem, then the j^{th} sample probably belongs to class 2.

It follows naturally that, in case of 2-class classification problems, this strategy can be employed for attempting to perform completely exact classifications. We will show some experiments in next section where good-quality predictions have been performed by using this strategy.

5. COMPUTATIONAL EXPERIMENTS

We present in this section some experiments related to the dataset of wine fermentations presented in Section 2 and that contains 24 fermentations of Cabernet sauvignon from a winery in Chile's Maipo Valley (Urtubia, Perez-Correa, Meurens, 2004). We select the features that are relevant for discriminating between normal and problematic fermentations by finding consistent biclusterings of the training set. This is done by solving a combinatorial optimization with a VNS-based heuristic recently proposed in (Mucherino, Cafieri, 2010). Then, the classifications of the samples of the testing set are performed by using the obtained biclustering. We also consider the strategy we presented in Section 4 for performing predictions without any misclassifications.

Table 1 shows some experiments in which the combinatorial optimization problem has been solved in order to find α-consistent biclusterings of A. $f(x)$ is the objective function of this optimization problem, which is a counter a selected features, that must be maximized. As expected, $f(x)$ decreases when the parameter α increases, because features subject to a noise or to an error that is larger than α are supposed to be removed from the set. *err* is the number of misclassifications on the testing set, when its samples are classified accordingly to the found α-consistent biclusterings. The biclustering with the largest α value is able to predict correctly 4 out of 8 samples. The corresponding biclustering related to the testing set is in fact not consistent.

Table 1: α-consistent biclusterings obtained from our database of wine fermentations

α	$f(x)$	*err*
0.00	448	5
1.00	445	5
1.50	442	4
1.70	440	4
2.00	438	4
2.10	431	4
2.20	431	4
2.30	402	4

In Table 2 we instead report some experiments that we performed for finding β-consistent biclusterings of A. As before, $f(x)$ becomes smaller and smaller when larger values for the parameter β are considered. In these experiments, we are able to discard many features that are potentially wrong or noisy, and we are able to define a β-consistent biclustering formed by 141 features only, that are supposed to be the most relevant for classification purposes. However, even in this case, not all samples of the testing set are correctly classified: 3 out of 8 are assigned to the wrong class.

For each of the 8 samples of the testing set, we also constructed the 8 matrices A_j, obtained by adding a column representing the j^{th} sample of the testing set to the training set A. Only the 141 features selected in the β-consistent biclusterings with $\beta = 1.80$ are considered.

As expected, the 5 matrices A_j that are related to the 5 samples that are correctly classified are all consistent. As a consequence, in case the classification of the samples of the testing set is not known, this information on the consistency of A_j can be exploited for validating the obtained predictions. The other 3 matrices are instead not consistent. We therefore assigned to the corresponding samples a classification which is the opposite of the one the supervised technique is able to provide. For two of the three matrices A_j, only one further feature has been removed in order to have the consistency of the biclustering associated to A_j. For one matrix, two features have been removed. For this reason, the actual classification of these three samples is probably the opposite of the one the procedure suggests. We verified the correctness of these results by removing these new features from the original training set A and by checking the consistency of the corresponding biclustering. All predictions in these experiments are correct.

Table 2: β-consistent biclusterings obtained from our database of wine fermentations

β	$f(x)$	*err*
1.00	448	5
1.20	397	5
1.30	340	5
1.40	281	4
1.50	262	3
1.60	211	3
1.70	147	3
1.80	141	3

6. CONCLUSIONS

We presented an analysis on a dataset of wine fermentations related to the harvest of 2002 of a Chilean winary. We considered a supervised biclustering technique that is based on the idea of constructing

consistent biclusterings in order to select the important features of a training set, and to exploit the obtained information for performing classifications for the samples of the testing set. We proposed in this paper a strategy for validating the results of the predictions, which may allow performing predictions with no mistakes on testing sets.

Future works will be mainly performed in the following two directions. First, larger datasets of wine fermentations need to be considered for obtaining better results. The fact that the considered testing set contains information which is not included in the training set suggests that it does not contain all necessary information for a correct definition of the biclusterings. Since industrial data are usually difficult to obtain, one possibility is to produce these data in laboratory, where small quantities of wine are fermented into a controlled environment. Moreover, we also plan to work on the formalization of the strategy that we proposed in this paper for validating the obtained classifications.

REFERENCES

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

Kundakcioglu, O.E., Pardalos, P.M., 2009. The Complexity of Feature Selection for Consistent Biclustering. In: S. Butenko, P.M. Pardalos, W.A. Chaovalitwongse (Eds.), *Clustering Challenges in Biological Networks*, World Scientific Publishing, 2009.

Madeira, S.C., Oliveira, A.L., 2004. *Biclustering Algorithms for Biological Data Analysis: a Survey*, IEEE Transactions on Computational Biology and Bioinformatics 1 (1), 24–44.

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

Mucherino, A., Papajorgji, P., Pardalos, P.M., 2009. *Data Mining in Agriculture*, Springer Optimization and Its Applications.

Mucherino, A., Urtubia, A., 2010. *Consistent Biclustering and Applications to Agriculture*, IbaI Conference Proceedings, Proceedings of the Industrial Conference on Data Mining (ICDM10), Workshop "Data Mining in Agriculture" (DMA10), Berlin, Germany, 105-113.

Nahapatyan, A., Busygin, S., Pardalos, P.M., 2008. An Improved Heuristic for Consistent Biclustering Problems. In: R.P. Mondaini and P.M. Pardalos (Eds.), *Mathematical Modelling of Biosystems*, Applied Optimization 102, 185–198.

Urtubia, A., Perez-Correa, J.R., Soto, A. Pszczolkowski, P., 2007. *Using Data Mining Techniques to Predict Industrial Wine Problem Fermentations*, Food Control 18, 1512–1517.