Recent Developments in Data Mining and Agriculture

A. Mucherino[1] and G. Ruß[2]

[1] CERFACS, Toulouse, mucherino@cerfacs.fr
[2] Otto-von-Guericke-Universität Magdeburg, russ@iws.cs.uni-magdeburg.de

Abstract. This survey covers some very recent applications of data mining techniques in the field of agriculture. This is an emerging research field that is experiencing a constant development. In this paper, we first present two applications in this field in details; in particular, we consider the problem of discovering problematic wine fermentations at the early stages of the process, and the problem of predicting yield production by using sensor data information. Secondly, we briefly describe other problems in the field for which we found very recent contributions in the scientific literature.

1 Introduction

Two years ago, one of the authors of this survey co-authored a book named "Data Mining in Agriculture" [22]. The book gives a wide overview of recent data mining techniques, and it also presents several applications in the field of agriculture, as well as in other related fields, such as biology. This is the first book completely devoted to this new and emerging research field. In fact, during the preparation of the book, applications of data mining in agriculture were not so common yet, and the task of finding and selecting references to be included in the book was not trivial to perform. Nowadays, after a couple of years, it seems this tendency has changed.

This survey covers the most recent applications in the field of data mining and agriculture, that have mainly been published in the period 2009–2011. Naturally, this survey is not meant to be exhaustive. We will give particular attention to two recent works in which the authors of this paper are directly involved, and then we will briefly mention some other applications that looked to us to be the most interesting to report.

The survey is organized as follows. In Section 2, we will present an analysis performed on datasets of wine fermentations with the aim of predicting problematic fermentations at the early stages of the process. Clustering and *supervised* biclustering techniques are employed for finding solutions to this problem. In Section 3, we will consider the problem of predicting yield production, in which state-of-the-art GPS technologies are employed in connection with site-specific and sensor-based treatments of crops. Various data mining techniques have been in this case tested for performing the predictions. In Section 4, a quick list of

other very recent applications in agriculture will be presented. Some final remarks will be given in Section 5.

2 Studying wine fermentations

Wine is widely produced all over the world. There exist different types of wine, which depend by different factors, and especially by the origin of the grapes that are employed in the production. A common point for all wines is the fermentation process, in which the sugar contained in the grapes is transformed in alcohol. This is a very delicate process. When producing wine industrially, indeed, large quantities of wine may get spoiled because of a problematic fermentation process, causing losses to the industry. In order to overcome to this issue, a prediction of the problematic wine fermentations could be attempted, so that an enologist can interfere with the process in time for guaranteeing a good fermentation.

In order to monitor wine fermentation processes, metabolites such as, for example, glucose, fructose, organic acids, glycerol and ethanol can be measured, and the data obtained during the fermentation process can be analyzed in order to obtain useful information. However, analyses are usually limited to data that are obtained within the first 3 days of fermentation. Naturally, this is done in order to learn about a possible problematic fermentation at the beginning of the process. Fermentations can be divided in 3 classes: the first class contains *normal* fermentations, while the second and the third one contain the problematic ones. In particular, the second class contains fermentations which are *slow*, in the sense that they can bring the wine to the end of the production, but in an amount of time which is longer than usual. Finally, the third class contains *stuck* fermentations, i.e. fermentations that stop at a certain moment and they are not able to give the final product.

Since 2004, a group of Chilean researchers are attempting the solution of this problem by using clustering techniques [31–33]. They consider a dataset containing 24 industrial vinifications of *cabernet sauvignon*, which are represented by several measurements performed during different fermentation processes. As a consequence, the dataset is composed by 22000 data points, each one representing a single measurement, that provides the levels of 30 chemical compounds involved in the fermentation.

Given a certain time t during the fermentation processes, measurements taken at time t can be grouped together in order to form clusters. A clustering technique might indeed define clusters that are related to normal or problematic fermentations by exploiting the inherent characteristics of the data. Naturally, due in large part to the time-variable nature of the fermentation process, fermentations can be assigned to different clusters for a different t. For this reason, a group of clusters can actually be defined for each fermentation. Fermentations that share the same group most likely share the same kind of characteristics. Depending on the percentage of normal, slow and stuck fermentations that are contained in the found groups of clusters, a score can be assigned to any other fermentation that happen to be in the same group and for which a classification

is not known. In these studies, the k-means algorithm [12] was employed for finding clusters of data points, where the number of clusters k was arbitrarily set to 5.

This technique is able to provide the enologist with a sort of *score* for each new fermentation that gives the probability for the fermentation to be problematic or not. However, no information regarding the compounds that causes the slow or the stuck fermentations are given, and this might be an important additional information in order to find the best way to interfere with the process. Therefore, more recently, supervised biclustering techniques have been applied to the same dataset of wine fermentations. This technique can simultaneously solve two data mining problems. First, it is able to select the features, the compound measurements, that are actually relevant in the fermentation process, so that useless data can be discarded, and compounds that may cause problematic fermentations can be identified. Second, the information that is acquired by finding biclusterings of the dataset can be exploited for performing classifications of new fermentations. Therefore, we can basically perform *feature selections* and *supervised classifications* at the same time by using this technique.

These studies have recently been published in [20, 21]. Each fermentation is here represented by a sample containing all compound measurements taken from the same fermentation process at different times. Samples are organized on the columns of a matrix A, and therefore measurements of the same compound taken from different fermentations, but at the same time t, can be found on the rows of A. For each compound and each time t, there is a specific feature in A. A *bicluster* is a submatrix of A defined by a subset of samples and a subset of features contained in A. As a consequence, a *biclustering* of A is a partition of A in disjoint biclusters, whose rows and columns cover the ones in A, and therefore it gives a relation between samples and features in A [3].

In order to select the pertinent features in A and to perform good classifications on new fermentations, it is required that the biclustering for A we find satisfies a property called *consistency*. If the biclustering is consistent, the samples and the features in the biclustering are strongly related to each other, so that the classification of its samples can be correctly obtained from the classification of its features, and vice versa. In order to find a consistent biclustering, a fractional optimization problem with binary variables can be defined, whose aim is to select the features that are actually relevant for the representation of the sample. This optimization is NP-hard [14]. A heuristic algorithm [19] can be used for the solution of this problem.

This biclustering technique was able to find some interesting information regarding the compounds that are monitored during the fermentation process [20]. For example, among the organic acids, the features related to lactic, malic, succinic, and tartaric acids are always preserved during the feature selection. Moreover, all the features related to each of these organic acids are assigned to only one bicluster, showing 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 other fermentations with

high levels of lactic acid are most likely going to get stuck as well. Moreover, the information on the levels of lactic acid seem to be very relevant starting from the first hours of the fermentation process, so that a prediction can be attempting at the very beginning of the process.

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 identify the features that are never selected in the biclusterings. 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 same compounds can be selected or discarded at different times t with irregular patterns. This behavior is different from the one obtained for sugar levels, for example, because sugar levels start to give relevant information only after some time from the beginning of the process (in fact, features related to sugar levels are always discarded when t is small). Deeper analysis are needed for understanding if compounds showing these irregular patterns can actually be important for the classification of the fermentations or not. Ongoing works are currently being performed in this direction.

The obtained biclusterings can be then exploited for performing supervised classifications of unknown classifications [21]. In order to verify the quality of the predictions, the dataset A can be divided in training and testing set: the training set can be used for performing the feature selection and for identifying consistent biclusterings, which can be successively used for predicting the classification of the samples in the testing set. The basic idea is to exploit the consistency of the biclustering for finding the classification of the samples of the testing set from the classification of its features (which is known because training and testing set have the same features). The technique is able to perform good-quality predictions of problematic fermentations.

3 Predicting yield production

Yield prediction is a very important agricultural problem. Any farmer would like, in fact, to know, as soon as possible, how much yield he can expect. Attempts to solve this problem date back to the time when first farmers began to work soils in order to get profit. Since years, yield predictions have been performed by considering farmer's experience on particular fields and crops. However, this knowledge can also be obtained by exploiting information given by modern technologies, such as GPS. A multitude of sensor data can nowadays be relatively easily collected, so that farmers do not only harvest crops but also growing and growing amounts of data. These data are fine-scale, often highly correlated and carry spatial information.

The problem of predicting yield production can be solved by employing data mining techniques. Consider that sensor data are available for some time back to the past, where the corresponding yield productions have been recorded. All this information form a training set of data which can be exploited to learn how to classify future yield productions, once new sensor data are available. There are different data mining techniques that can be used for this purpose. In [25], for example, two different neural networks are considered, one network with a multi-layer perception, another one with a radial basis function, as well as a support vector regression and a decision regression tree. A comparison of these four techniques showed that the support vector regression technique is the most suitable for this kind of problem.

Moreover, in order to improve the quality of the predictions, the concept of *spatial autocorrelation* has more recently been considered in [26]. When considering the data mining techniques mentioned above, it is implicitly supposed that the data are not correlated. However, with the given geo-tagged data records at hand, this is clearly not the case, due to their (natural) spatial autocorrelation. Therefore, the spatial relationships between data records should be taken into account. Spatial autocorrelation [7] is the correlation among values of a single variable strictly attributable to the proximity of those values in geographic space, introducing a deviation from the independent observations assumption of classical statistics. For the data sets related to this problem, each of the attributes exhibits spatial autocorrelation. Usually, it is known from the data origin whether spatial autocorrelation exists.

In non-spatial models, data records which appear in the training set are not supposed to appear in the test set during a cross-validation learning setup. Classical sampling methods do not take spatial neighborhoods of data records into account. Therefore, the above assumption may be rendered invalid when using non-spatial models on spatial data. This inevitably leads to overfitting and underestimates the true prediction error of the regression model. Therefore, the main issue is to avoid having neighboring or the same samples in training and testing data subsets during a cross-validation learning approach. The basic idea is to apply changes to the resampling method and keep the regression modeling techniques as-is. The resulting procedure can be seen as a spatial cross-validation technique.

In general, when considering the k-fold cross-validation technique, the original dataset can be divided in three parts: a training set, a validation set and a test set. Setting k equal to 10 or 20 is generally considered to be appropriate to remove bias. The regression model is trained on the training set until the prediction error on the validation set starts to rise. Once this happens, the training process is stopped and the error on the test set is reported for this fold.

In spatial data, due to spatial autocorrelation, almost identical data records may end up in training, validation and test sets. In essence, the model overfits the training data and returns an overoptimistic (biased) estimation of the prediction error. Therefore, one possible solution might be to ensure that only a very small number (if any) of neighboring and therefore similar samples end up in training

and test subsets. This can be achieved by adapting the sampling procedure for spatial data. Once this issue is accommodated, the cross-validation procedure can continue in the usual way.

A spatial clustering procedure can be employed to subdivide the fields into spatially disjunct clusters or zones. The clustering algorithm can then be run on the data records' spatial map, using the data records' longitude and latitude. Depending on the clustering algorithm parameters, this results in a tessellation map which does not consider any of the attributes, but only the spatial neighborhood between data records. In analogy to the non-spatial regression treatment of these data records, a spatially aware cross-validation regression problem can therefore be handled using the k resulting zones of the clustering algorithm as an input for k-fold cross-validation. This ensures that the training set has only a small amount of spatial autocorrelation with the test set. Standard models can be used straightforwardly, without requiring changes to the models themselves.

First computational experiments can be found in [26], which show that it is actually important to closely consider spatial relationships inherent in the data sets in this kind of data mining problems. This work proves that, if spatial autocorrelation exists, standard regression models should be adapted to the spatial case.

4 Other recent works

We mention in this section some other recent interesting works in the field of data mining and agriculture, mainly published between 2009 and 2011. We begin with some other works related to the production of wine, which has been the focus of Section 2, where data mining approaches are employed for the prediction of problematic wine fermentations. The main aim of this work is to discover in advance fermentations that are going to be slow or stagnant, and to interfere with the process in order to guaranteeing a good fermentation. Other recent studies also concern the taste of the wine that is produced. In [4, 24], for example, data mining techniques are employed in order to predict the taste of wine. This is done by creating a training set in which a classification of each sample (wine) is assigned by traditional wine tasters, that generally analyze some subjective parameters such as color, foam, flavor and savour of the wine. Once the classification task has been learned by exploiting the training set, data mining techniques are then supposed to substitute traditional wine tasters. Wine tastes are also analyzed in [28] in relation to seasonal climate effects.

In [22], many applications in the field of agriculture have been presented in details. It is interesting to remark that some of the papers related to such applications have been cited various times meanwhile, and some important developments have been presented. Starting from [16], for example, the problem of recognizing and grading fruits by using automatic data mining techniques has been recently also studied in [1, 5, 30]. In [27], a non-destructive technique has been discussed for identifying defects in apples, and, more recently, studies have been presented where new developments in this research direction can be

found [2, 10, 11]. Finally, starting from the works previously presented in [18], new studies on the detection and the analysis of sounds issued by animals have recently been proposed in [8].

Recent works in the field especially regard China. This is the most populous country in the world, and it is also the major emitter of greenhouse gases. For this reason, the relation between climate changes, water resources and agriculture in China has been deeply analyzed in [23]. Pig industry also plays a very important role in adjusting the agricultural structure of China. Since pig prices are likely to fluctuate very violently, the study presented in [6] has as main aim the prediction of the price of pigs in the Chinese market.

Other recent applications include an automatic classification for flower species, where a k-nearest neighbor classifier is employed [9]. Artificial neural networks are instead used for predicting the total necessary power of agricultural machinery [15], whereas the estimation of soil properties and the classification of soil types is performed by employing support vector machines [13]. Finally, the prediction of foodborne disease outbreaks and the forecast of water consumption in agriculture are studied, respectively, in [29] and [17].

5 Conclusions

This review presents a quick update with respect to the state-of-the-art in the field of data mining and agriculture given in [22]. We mainly focus our attention on two particular problems. The first one is the problem of identifying problematic wine fermentations at the early stages of the process. In [22], a data mining approach to this problem has been discussed where the k-means algorithm was used. We described the recent developments on this problem, and in particular new studies where biclustering techniques are employed for identifying the compounds of wine that are most likely the cause of problematic fermentations. The second problem we consider is the one of predicting yield production. First approaches to this problem were based on standard data mining techniques, such as support vector regression and artificial neural networks. Recent works showed how to improve the quality of the classifications by employing the concept of spatial autocorrelation. Other recent applications in the field, which have mainly been published in the period 2009–2011, are also quickly reviewed.

References

1. S. Arivazhagan, R.N. Shebiah, S.S. Nidhyanandhan, L. Ganesan, *Fruit Recognition using Color and Texture Features*, Journal of Emerging Trends in Computing and Information Sciences **1**(2), 90–94, 2010.
2. P. Baranowski, W. Mazurek, *Detection of Physiological Disorders and Mechanical Defects in Apples using Thermography*, International Agrophysics **23**, 9–17, 2009.
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. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, *Modeling Wine Preferences by Data Mining from Physicochemical Properties*, Decision Support Systems **47**(4), 547–553, 2009.
5. S. Cubero, N. Aleixos, E. Moltó, J. Gómez-Sanchis, J. Blasco, *Advances in Machine Vision Applications for Automatic Inspection and Quality Evaluation of Fruits and Vegetables*, Food and Bioprocess Technology **4**(4), 487–504, 2011.
6. L. Ding, J. Meng, Z. Yang, *An Early Warning System of Pork Price in China Based on Decision Tree*, IEEE Conference Proceedings, International Conference on E-Product E-Service and E-Entertainment (ICEEE), Henan, China, 1–6, 2010.
7. D.A. Griffith, *Spatial Autocorrelation and Spatial Filtering*, Advances in Spatial Science Series, Springer, New York, 2003.
8. M. Guarino, P. Jans, A. Costa, J-M. Aerts, D. Berckmans, *Field Test of Algorithm for Automatic Cough Detection in Pig Houses*, Computers and Electronics in Agriculture **62**(1), 22–28, 2008.
9. D.S. Guru, Y.H. Sharath, S. Manjunath, *Texture Features and KNN in Classification of Flower Images*, International Journal of Computer Applications **1**, Special Issue "Recent Trends in Image Processing and Pattern Recognition", 21–29, 2010.
10. R.P. Haff, *Real-Time Correction of Distortion in X-ray Images of Cylindrical or Spherical Objects and its Application to Agricultural Commodities*, Transactions of the American Society of Agricultural and Biological Engineers **51**(1), 341–349, 2007.
11. R.P. Haff, N. Toyofuku, *X-ray Detection of Defects and Contaminants in the Food Industry*, Sensing and Instrumentation for Food Quality and Safety **2**(4), 262–273, 2008.
12. J. Hartigan, *Clustering Algorithms*, John Wiles & Sons, New York, 1975.
13. M. Kovacevic, B. Bajat, B. Gajic, *Soil Type Classification and Estimation of Soil Properties using Support Vector Machines*, Geoderma **154**(3–4), 340–347, 2010.
14. 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.
15. Q. Lai, H. Yu, H. Chen, Y. Sun, *Prediction of Total Power of Agricultural Machinery Using Artificial Neural Networks*, IEEE Conference Proceedings, International Conference on Computing, Control and Industrial Engineering, volume 2, Wuhan, China, 394–396, 2010.
16. V. Leemans, M.F. Destain, *A Real Time Grading Method of Apples based on Features Extracted from Defects*, Journal of Food Engineering **61**, 83-89, 2004.
17. S. Lu, Z-J. Cai, X-B. Zhang, *Forecasting Agriculture Water Consumption based on PSO and SVM*, IEEE Conference Proceedings, 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, China, 147–150, 2009.
18. D. Moshou, A. Chedad, A. Van Hirtum, J. De Baerdemaeker, D. Berckmans, H. Ramon, *Neural Recognition System for Swine Cough*, Mathematics and Computers in Simulation **56**, 475-487, 2001.
19. A. Mucherino, S. Cafieri, *A New Heuristic for Feature Selection by Consistent Biclustering*, arXiv e-print, arXiv:1003.3279v1, March 2010.
20. A. Mucherino, A. Urtubia, *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, 2010.
21. A. Mucherino, A. Urtubia, *Feature Selection for Datasets of Wine Fermentations*, I3M Conference Proceedings, 10^{th} International Conference on Modeling and Applied Simulation (MAS11), Rome, Italy, September 2011.

22. A. Mucherino, P. Papajorgji, P.M. Pardalos, *Data Mining in Agriculture*, Springer, 2009.
23. S. Piao, P. Ciais, Y. Huang, Z. Shen, S. Peng, J. Li, L. Zhou, H. Liu, Y. Ma, Y. Ding, P. Friedlingstein, C. Liu, K. Tan, Y. Yu, T. Zhang, J. Fang, *The Impacts of Climate Change on Water Resources and Agriculture in China*, Nature **467**, 43–51, 2010.
24. J. Ribeiro, J. Neves, J. Sanchez, M. Delgado, J. Machado, P. Novais, *Wine Vinification Prediction using Data Mining Tools*, Conference Proceedings, Computing and Computational Intelligence, Tbilisi, Republic of Georgia, 78–85, 2009.
25. G. Ruß, *Data Mining of Agricultural Yield Data: A Comparison of Regression Models*, Conference Proceedings "Advances in Data Mining – Applications and Theoretical Aspects", P. Perner (Ed.), Lecture Notes in Artificial Intelligence **6171**, Berlin, Heidelberg, 24–37, Springer, 2009.
26. G. Ruß, A. Brenning, *Data Mining in Precision Agriculture: Management of Spatial Information*, Conference Proceedings, Computational Intelligence for Knowledge-Based Systems Design, E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.), Lecture Notes in Artificial Intelligence **6178**, Berlin, Heidelberg, 350–359, Springer, 2010.
27. T.F. Schatzki, R.P. Haff, R. Young, I. Can, L-C. Le, N. Toyofuku, *Defect Detection in Apples by Means of X-ray Imaging*, Transactions of the American Society of Agricultural Engineers **40**(5), 1407-1415, 1997.
28. S. Shanmuganathan, P. Sallis, A. Narayanan, *Data Mining Techniques for Modelling Seasonal Climate Effects on Grapevine Yield and Wine Quality*, IEEE Conference Proceedings, Second International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN10), Liverpool, UK, 84–89, 2010.
29. M. Thakura, S. Olafssonb, J-S. Leeb, C.R. Hurburgha, *Data Mining for Recognizing Patterns in Foodborne Disease Outbreaks*, Journal of Food Engineering **97**(2), 213–227, 2010.
30. D. Unay, B. Gosselin, O. Kleynen, V. Leemans, M-F. Destain, O. Debeir, *Automatic Grading of Bi-colored Apples by Multispectral Machine Vision*, Computers and Electronics in Agriculture **75**(1), 204–212, 2011.
31. A. Urtubia, J.R. Perez-Correa, M. Meurens, E. Agosin, *Monitoring Large Scale Wine Fermentations with Infrared Spectroscopy*, Talanta **64** (3), 778-784, 2004.
32. 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.
33. 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.