Abstract:
This paper shows a brief thought of a percentage of the broadly utilized data mining systems over agri-business information sets. It is a contemporary system to discover the arrangement over the customary and traditional technique. It is a new concept as for data mining in the field of farming. Data mining is a procedure of separating/comparing the important information from the extensive database by utilizing diverse tools and methods. Data mining is the procedure of assessing information from various standpoints, extracting only the vital data and outlining it into helpful data form. This investigation performed on datasets of wine fermentations with the point of foreseeing issues related to the fermentation at the early phases of the procedure. We have tried to analyze different data mining procedures related to the agriculture field for achieving the expected success.

Keywords: Data mining techniques in agriculture, Prediction of problematic wine fermentations, Applications.

[1] INTRODUCTION
i. DATA MINING

Data Mining is the process of extracting useful and important information from large sets of data. Data mining in agriculture field is a relatively novel research field. In this paper describe an overview of Data Mining techniques applied to agricultural and their applications to agricultural related areas.

Data mining is that the method of analyzing knowledge from completely different views and summarizing it into helpful data that may be won’t to increase revenue, cuts costs, or both. Data processing computer code is one in every of the variety of analytical tools for analyzing knowledge. It permits users to research knowledge from many alternative dimensions or angles, reason it, and summarize the relationships known. Technically, {data mining and data method}
is that the process of finding correlations or patterns among dozens of fields in giant relative databases.

Data mining, or data discovery, is that the computer assisted method of analyzing monumental sets of information so extracting which means of the info. Data processing tools predict behaviours and future trends, permitting businesses to form proactive, knowledge driven selections. Data processing tools will answer business queries that historically were too moment overwhelming to resolve.

ii. AGRICULTURE

Agriculture is that the cultivation of animals, plants, fungi, and different life forms for food, fibre, and bio fuel, healthful and different merchandise accustomed sustain and enhance human life. Agriculture was the key development within the rise of inactive human civilization, whereby farming of domesticated species created food surpluses that nurtured the event of civilization. The study of agriculture is understood as agricultural science. The history of agriculture dates back thousands of years, and its development has been driven and outlined by greatly completely different climates, cultures, and technologies. Within the civilized world, industrial agriculture supported large-scale monoculture farming has become the dominant agricultural methodology.

Data mining in agriculture is a very recent research topic. It consists in the application of data mining techniques to agriculture. Recent technologies use different units that are able to offer lots of data on agricultural related activities, which might be analyzed so as to seek out necessary data.

[2] DATAMINING TECHNIQUES USED IN AGRICULTURE

Data mining in agriculture is a very recent research topic. It consists in the application of data mining techniques to agriculture. This data mining techniques used in the field of agriculture is helpful in prediction of problems, disease detection, optimizing the pesticide and so on. Recent technologies are nowadays able to provide a lot of information on agricultural-related activities, which can then be analyzed in order to find important information and to collect relevant information. This data mining techniques are used for disease detection, pattern recognition by using multiple application.

Data mining is about to identify the similarities between searching the valuable business information from the large database systems such as finding linked products in gigabytes of store scanner data or the mining a mountain for a vein of valuable dataset. Data mining can be done on a database whose size and quality are sufficient. The technology of data mining can generate new business opportunities by providing these capabilities:

- Automated prediction and analysis of various trends and behaviours - Data mining itself automate the process by obtaining the predictive information from large databases. It first setup the questions and then provides the relative solutions. A typical example of such a predictive system is in the marketing field. Data mining uses data on the historical promotional mailings to capture the targets effectively so that the maximum return from market will be achieved.

- Another application of data mining is automated discovery of historical patterns dynamically. The presented Data mining system is able to sweep over the databases to identify the hidden
patterns. One of such example of pattern discovery is the analysis of retail sales data to identify the seemingly unrelated products so that the effective purchase can be done.

Data mining techniques are able to get the benefits of automation on existing software and hardware platforms that can be implemented on new systems which can be upgraded and new products can be developed. When data mining tools are defined on high performance parallel systems, they can be analyzed with massive databases in minutes. High speed processing and accurate outcome from the system makes it possible for users to analyze large set of information. Larger databases, in turn, gives more improved predictions.

[3] PREDICTION OF PROBLEMATIC 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. They consider a dataset containing 24 industrial verifications 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.

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 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.

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.

In order to verify the quality of the predictions, the dataset 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.

[4] 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 data 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.

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.

Moreover, in order to improve the quality of the predictions, the concept of spatial autocorrelation has more recently been considered in. 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 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.
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 neighbourhoods 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 neighbouring 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 these regression modelling 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 into 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 neighbouring 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 neighbourhood 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, 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.

[5] APPLICATIONS OF DATA MINING TECHNIQUES IN AGRICULTURE

i. Sorting apples by water cores:

Before going to market, apples are checked and the ones demonstrating a few deformities are evacuated. In any case, there are additionally undetectable imperfections that can ruin the apple flavour and look. A sample of undetectable deformity is the watercore. This is an internal apple issue that can influence the life span of the natural product. Apples with slight or gentle watercores are sweeter, yet apples with moderate to disjoint level of watercore can’t be put away for any period of time. In addition, a couple natural products with serious watercore could ruin an entire bunch of apples. Consequently, a computational framework is under study which takes X-beam photos of the natural product while they keep running on transport lines, and which
is additionally ready to break down (by information mining strategies) the taken pictures and gauge the likelihood that the organic product contains watercores.

**ii. Optimizing pesticide usage by data mining:**

Recent studies by agriculture researchers showed that attempts of cotton crop yield maximization through pro-pesticide state policies have led to a dangerously high pesticide usage.

**iii. Explaining pesticide abuse by data mining:**

To screen cotton development, distinctive government offices and organizations have been recording nuisance scouting, farming and metrological information for quite a long time. Coarse evaluations of simply the cotton bother scouting information recorded stands at around 1.5 million records, and developing. The essential agro-met information recorded has never been digitized, incorporated or institutionalized to give a complete picture, and henceforth can't bolster choice making, along these lines requiring an Agriculture Data Warehouse. Making a novel Pilot Agriculture Extension Data Warehouse finished by investigation questioning and information mining some fascinating disclosures were made, for example, pesticides splashed at the wrong time, wrong pesticides utilized for the right reasons and fleeting relationship between pesticide utilization and day of the week.

[6] **CONCLUSION**

Agricultural organizations and their management try every day to find information in large databases for business decision making. Often the case is that the solution for their problems was within their reach and the competition has already used this information. Data mining, through better management and data analysis, can assist agricultural organizations to achieve greater profit. Therefore it is crucial that managers of agricultural organizations get to learn about the idea and techniques, because the amount of available information is sure to grow in the future, and it will not become clearer and easier to understand and make decisions. Understanding of the processes which are carried out and decisions being made in agricultural organizations is enabled through data mining. By the use of data mining technique acquired knowledge can be used to make successful decisions which will advance the success of the agricultural organization on the market. Data mining once started, presents an endless cycle of acquiring knowledge. For organizations it represents on of the key points to create a business strategy on. Great efforts are invested in finding a more successful application of data mining in agricultural organizations.

**REFERENCES**

[1] CERFACS, Toulouse, mucherino@cerfacs.fr, Otto-von-Guericke-Universit¨at at Magdeburg Recent Developments in Data Mining and Agriculture

