ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 3, pp: (158-162), Month: July - September 2014, Available at: www.researchpublish.com

Application of Data Mining In Agriculture

P. Grace Sharon

Student, Department of Information Technology, Saveetha School of Engineering,
Saveetha University, Chennai– 602 105, India

Abstract: **Data mining is the study of the Knowledge Discovery in Databases process which is known as KDD. It is an interdisciplinary subfield of computer science, the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Data mining in agriculture is a very recent research topic. It consists in the application of data mining techniques to agriculture. Recent technologies are able to provide a lot of information on agricultural-related activities, which can then be analyzed in order to find important information. A related equivalent term is precision agriculture. In this paper the applications such as 1. Prediction of wine fermentation problems and 2.grading of apples are discussed.**

Keywords: **Agriculture, Data mining.**

1. INTRODUCTION

The k-means algorithm and all its variants, and the fuzzy approach have been applied to a wide variety of real-life problems. The k-means and fuzzy c-means algorithm are for instance used for analyzing and categorizing gene expression data, in order to analyze and presume the function of unknown genes. The k-means algorithm has been applied to solve the problem of segmenting images with smooth surfaces the genetic k-means algorithm has been applied for compressing images; the y-means algorithm has been developed for monitoring instructions in computer systems; the fuzzy c-means has been applied for detecting crime hot-spots or geographic areas of elevated criminal activity.

In the field dog agriculture, the k-means algorithm has been applied for forecasting pollution in the atmosphere; soil classifications using GPS-based technologies; classification of plant, soil, and residue regions of interest by color images; predicting wine fermentation problems; grading apples before marketing; monitoring water quality changes; detecting weeds in precision agriculture. The two applications in agriculture are discussed in detail in the following. The problem of predicting the fermentation process of wine and classifying it as good or bad is presented in section 1.1. the problem of classifying apples on basis of their grade is discussed in section 2.1.

1.1 PREDICTION OF WINE FERMENTATION PROBLEMS

Problems occurring during the fermentation process can impact the productivity of wine-related industries and the quality of wine. The fermentation process of wine is too slow or it can even become stagnant. Predicting how good the fermentation process is going to be may help enologists (wine specialists) who can then take suitable steps to make corrections when necessary and to ensure that take suitable steps to make corrections when necessary and to ensure that the fermentation process concludes smoothly and successfully. On order to monitor the wine fermentation process, metabolites such as glucose, fructose, organic acids, glycerol and ethanol can be measured, and the data obtained during the entire fermentation process can be analyzed in order to obtain useful information. Data mining techniques can help extract this information from large databases, which may be able to predict the fermentation process. In the work which is the focus of this section, a k-means algorithm has been applied for exploring data accumulated from measurements sampled regularly of 24 industrial vinifications of cabernet sauvignon. Data measured during the first three days of fermentation has been compared to those obtained during the whole fermentation process. Information on the behavior of the fermentation during the first three days can provide important clues about the final classification.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 3, pp: (158-162), Month: July - September 2014, Available at: www.researchpublish.com

1.2. METHODOLOGY

The data come from a winery in Chile's Maipo Valley, and they are related to the 2002 harvest. Between 30 and 35 samples are taken per fermentation depending on the duration of a vinification. The levels of 29 compounds are analyzed. Among them, sugars are analyzed, such glucose and fructose, organic acids, such as the lactic and citric acids, nitrogen sources, such as alanine, arginine, leucine etc., and alcohols. The whole set of data consists in approximately 22,000 data points. The used compounds are actually 28 since taking glucose and fructose as a single variable which is known to be sugar is the same as considering the two sugars as independent variables. Four sets of data are defined in order to perform the analysis. Datasets A and B just consider 8 variables, including "sugars", alcohols and organic acids, whereas datasets E and F include all 28 components. The data contained in datasets A and E are related to the first three days of fermentation, whereas datasets B and F are related to data measured during the whole fermentation process.

Graphic representation of the considered databases

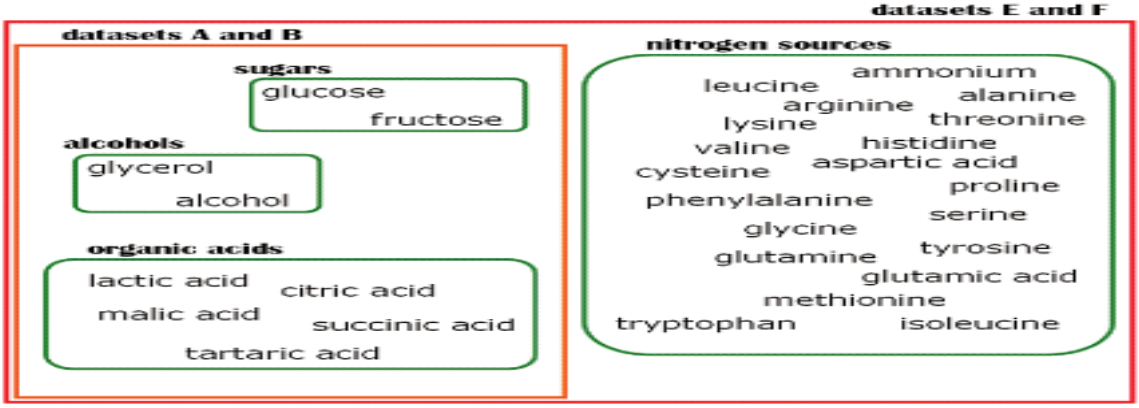

Fig.1 A graphical representation of the compounds considered in datasets A, B, E and F. A and E are related to data measured within the three days that the fermentation started; B and F and related to data measured during the whole fermentation process.

2. GRADING OF APPLES BY WATERCORES

X-ray radiograms were obtained for 400 to 700 each of four State of Washington apple cultivars ('Fuji', 'Granny Smith', 'Red' and 'Golden Delicious') and 79 'Braeburns' carrying assorted defects (bruises, senescence browning, rot, water core and insect damage). Radiograms of whole apples, most out of long term storage, were obtained with line scanning x-ray, suitable for real-time inspection, and with high resolution film, at two orientations, following which the apples were sliced and photographed. Apples were characterized as defective or not based on the appearance in these photos. Sets of x-ray images for a given cultivar/orientation (good and bad apples mixed randomly) were inspected by human observers and the recognition rates recorded. When still images were viewed on a computer screen, acceptable recognition (= 50% of defective apples recognized, = 5% of good apples classified defective) of images was obtained for senescence browning of 'Red Delicious', for water core and stem rot in 'Fuji' (requiring orientation), possibly for water core in 'Red Delicious', and for codling moth damage in the first four cultivars 8 to 19 days after larval entry. However, when images were scrolled across the screen at increasing rates, simulating a three-chain sorting line, recognition fell off to unacceptable levels at rates one half that corresponding to a commercial sorting line. This decrease is not unexpected in light of previous work in more structured situations. The implications for devising an apple sorting system based on x-ray are discussed.

Clustering techniques are used for finding suitable groupings of samples belonging to a given set of data. There is no knowledge a priori about these data. Therefore, such set of samples cannot be considered as a training set, and classification techniques cannot be used in this case. The *k*-means algorithm is one of the most popular algorithms for clustering. It is one of the most used algorithms for data mining, as it has been placed among the top 10 algorithms for data mining in.

The *k*-means algorithm partitions a set of data into a number *k* of disjoint clusters by looking for inherent patterns in the set. The parameter *k* is usually much smaller than the dimension of the set of samples, and, in general, it needs to have a predetermined value before using the algorithm. There are cases where the value of *k* can be derived from the problem studied. For instance, in the example of the blood test analysis (see Section 1.1), the aim is to distinguish between healthy and sick patients. Hence, two different clusters can be defined, and then $k = 2$. In other applications, however, the parameter *k* may not be defined as easily. In the example of separating good apples from bad ones, images of apples need to be analyzed. The set of apple images can be partitioned in different ways. One partition can be obtained by dividing apples into two clusters, one containing apples with defects and another one containing good apples. In this case $k = 2$. However, defective apples can be classified based on the degree of the defect. For instance, if the apples have a defect which is not very visible, then these apples could be sold with a lower price. Therefore, even defective apples can be grouped in different clusters. In this case, *k* shows the number of defects that are taken into consideration. When there is uncertainty on the value of the parameter *k*, a set of possible values is considered and the algorithm is carried out for each of the values. The best obtained partition in clusters can then be considered.

3. METHODOLOGY

There are machines able to take picture of the fruits while they are passing through them. Usually, fruits are placed on rolled which make the apples rotate on themselves and the pictures are taken from a camera located above. In this case, the parts of the fruit close to the points where the rotation axis acrosses its surface may not be observed. Hence, if some defect is there, it may not be identified, but this problem can be overcome by placing mirrors on each side of the rollers. More complex systems have also been developed, in which fruits are free to move on ropes while three cameras take pictures from different places, or where robot arms are used to manipulate the fruit. The system which uses robot arms was able to observe 80% of the fruit surface with four images, but it is quite slow, since it takes about 1 second for analyzing 4 fruits. Another important issue is the lighting system used. Commonly the images are monochrome images, but they can also be color images.

After the image or the images have been acquired from the apple, the segmentation process must be applied. The result of image segmentation is the division of such image in many regions, related for instances to different gray levels, that represent the background, the healthy tissue of the fruit, the calyx, the stem and possible defects. The contrast between the fruit and the background should be high to simplify the localization of the apple, even though calyx, stem ends and defects may have the same color of the image background. The hard task is how to separate the defects from the healthy tissue, the calyx and the stem. On monochrome images, the apple appears in light gray, the mean luminance of the fruit varies with its color and decreases from the center of the fruit to the boundaries. Defects are usually darker than the other regions, but their size and their shape can vary strongly.

Supervised or unsupervised techniques can be used to segment the obtained images. The supervised techniques tend to reproduce a pre-existent classification or segmentation and the unsupervised techniques produce segmentation on their own. For example, the neural networks have been used for classifying pixels into six classes including a class representing the fruit defect. The work which is the focus of the section is instead based on a k-means approach, which is an unsupervised technique, since it is able to partition the data without having any previous knowledge about them.

This approach is different from the others because it manages several images representing the whole surface of the apple at the same time. In other works, each image taken from the same fruit was treated separately and the fruit was classified according to the worst result of the set of representative images. The method discussed here combines instead the data extracted from the different images of a fruit moving on a machine in order to dispose information related to the whole surface of the fruit. The method is applied on Jonagold apples characterized by green which is ground color and red which is blush colors.

Images representing different regions of the fruit are analyzed and segmented. The regions issued from segmentation process including the defects, over-segmentation and calyx and stem ends are called blobs. These regions are characterized by using color or gray scale, position, shape and texture features. In total, 15 parameters are considered for characterizing a blob, five for the color, four for the shape, five for the texture and one for the position. The k-means

algorithm, the blob and fruit discriminant analysis are made off-line by the program, whereas the blobs and afterwards the fruits can be graded in-line by using the parameters of the discriminant analysis.

Once the clusters have been defined, apples are classified with a global correct classification rate of 73%. These results have been obtained by using a set containing 100 apples, i.e. 100 apples have been partitioned for obtaining the set of clusters successively used for classifying other unknown apples.

Fig.2 X-ray line scans of Red Delicious apples. Raw images are on the left. Computer enhanced zones of water core are depicted on the right (Shahin and Tollner, 1997)

4. CONCLUSION

This topic is a recent concept and has a very good beneficial towards the agriculture and data mining plays an important role where clustering and k-means algorithms are widely being used. Data mining and knowledge discovery techniques are relatively new to agricultural and environmental fields. Their use is associated and conditioned with the use of research operations sets of tools.

There are many research papers that show agricultural and environmental sciences are really beneficial from the use of mathematical tools and modern technology. It is important to note that the a number of published papers are purely research and have not yet been applied to be part of the set of tools farmers or practitioners use every day. As an example, the study that use a Artificial Neural Network (Aerts et al., 2004) describes the process of how a pig goes through the system designed to record pig coughs to discover whether the animal has health problems but it does not address the issue of scaling the proposed system to be applied to the entire herd.

Thus data mining and its techniques provide a very good benefits in the future technology.

ACKNOWLEDGMENT

The author would like to thank Mr. Fahad Iqbal for his guidance and support.

REFERENCES

[1] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.

[2] Mucherino, A.; Papajorgji, P.J.; Pardalos, P. (2009). *Data Mining in Agriculture*, *Springer*. http://postharvest.tfrec.wsu.edu/pages/J8I4D

[3] 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) Springer.

[4] T. F. Schatzki, R. P. Haff, R. Young, I. Can, L-C. Le, N. Toyofuku, the American Society of Agricultural and Biological Engineers, St. Joseph, Michigan, www.asabe.org, Transactions of the ASABE. VOL. 40(5):1407-1415. (doi:10.13031/2013.21367) @1997

[5] Data Mining techniques in agricultural and environmental sciences Altannar Chinchuluun, University of Florida, USA Petros Xanthopoulos, University of Florida, USA Vera Tomaino, University of Florida, USA and University Magna.