2010 First International Conference on Pervasive Computing, Signal Processing and Applications

The graphical feature extraction of star plot for wine quality classification

LI Jing
College of Science; College of Electrical Engineering
Yanshan University
Qinhuangdao, China
01016888@sina.com

WANG Jin-Jia, ZHANG Tao
College of Information Science and Engineer
Yanshan University
Qinhuangdao, China
wjj@ysu.edu.cn

MA Chong-Xiao
Determent of Machinery & Electron
Hebei Normal University of Science & Technology
Qinhuangdao, China

HONG Wen-Xue
College of Electrical Engineering
Yanshan University
Qinhuangdao, China

Abstract— **we propose a visualization method of evaluation of wine quality. The wine data are from the certification phase of the physicochemical analysis test. The data include the 11 input variables, an output variable which is the quality of wine. The data include 1599 samples of red wine and 4898 samples of white wine. Our method works better than the traditional neural networks and support vector machine method, and has visual advantages. Such model is useful to support the oenologist wine tasting evaluations and improve wine production. Furthermore, similar techniques can help in target marketing by modeling consumer tastes from niche markets.**

Keywords- visual evaluation; support vector machines; feature extraction; graphical representation of the multivariate data

I. INTRODUCTION

In recent years，wine industry has been developing in a high speed．However, the quality status of wine is not satisfied．For the commercial wine，the quality of some products are not up to the requirements of the national standards，the labels of some are not right，or the wine age of some are wrongly labeled．The wines have increased very much by from 1997 to 2010 [1]. To support its growth, the wine industry is investing in new technologies for both wine making and selling processes. Wine cortication and quality assessment are key elements within this context. Cortication prevents the illegal adulteration of wines and assures quality for the wine market. Quality evaluation is often part of the cortication process and can be used to improve wine making and to stratify wines such as premium brands.

Wine cortication is generally assessed by physicochemical and sensory tests. Physicochemical laboratory tests routinely used to characterize wine include determination of density, alcohol or pH values, while sensory tests rely mainly on human experts. It should be stressed that taste is the least understood of the human senses [3], thus wine classification is difficult task. Moreover, the relationships between the physicochemical and sensory analysis are complex and still not fully understood.

Pattern recognition technologies have made it possible to collect, store and process massive, often highly complex datasets. All this data hold valuable information such as trends and patterns, which can be used to improve decision making and optimize chances of success [4]. There are several pattern recognition algorithms, each one with its own advantages.

The one keystone in pattern recognition is that how the mathematic features are selected and extracted by the learning observations [5]. The issue of representation is an essential aspect of pattern recognition and is different from classification. It largely influences the success of the stages to come. This is a promising direction. Building proper representations has become an important issue in pattern recognition.

As we know, the graphical representation or graphical analysis for multidimensional data in multivariate analysis is a very useful tool [6]. One commonly used graphical representation form is the 'star plot', and one interesting graphical representation form is the 'Chernoff faces'. But the graphical representation for multidimensional data has a limited application in pattern recognition. WANG have done some work in the graphical representation for pattern recognition using star plot [7-8] and Chernoff faces [9-10]. Experiments with several standard benchmark data sets show the effectiveness of the new graphical features.

In this research，graphical feature were applied to determine wine quality. The new graphical features are evaluated by the various classifiers, which include linear classifier; k-nearest neighbor classifier, neural network classifier and Support vector classifier.

II. GRAPHICAL REPRESENTATION AND GRAPHICAL FEATURES

A. Graphical Representation

Given $\{x_j \in R^d, j = 1,2,\cdots N\}$ of the multi-dimension data, $x_j = \{x_{1j},\cdots,x_{ij},\cdots,x_{dj}\}$ represent each sample. In general, the each data dimension first is

978-0-7695-4180-8/10 $26.00 © 2010 IEEE
DOI 10.1109/PCSPA.2010.192

normalized to the predefine interval. The transform formula is

$$x_{ij}^{'} = a_j + (b_j - a_j)\frac{x_{ij} - x_{\min j}}{x_{\max j} - x_{\min j}}, \qquad (1)$$

$$\mathrm{j} = 1, \cdots, \mathrm{N}; \quad \mathrm{i} = 1, \cdots d$$

where $x_{ij}^{'}$ is the transformed data; x_{ij} is the i-th dimension of the j-th sample; x_{minj} and x_{maxj} represent the minimum and maximum of the i-th dimension; a_j and b_j represent the upper and lower limits of the predefine interval. Finally all $x_{ij}^{'}$ fall in $[a_j, b_j]$, which $a_j = 0$，$b_j = 1$.

The star plot is a simple means of multivariate visualization, which represents the value of an attribute through the length of lines radiating from the icon's center. Figure 1 displays star plots of the IRIS data. Each symbol displays all four variables. It is created by the Matlab function glyphplot(X), which creates a star plot from the multivariate data in the n-by-p matrix X. Rows of X correspond to observations, columns to variables. A star plot represents each observation as a "star" whose i-th spoke is proportional in length to the i-th coordinates of that observation. glyphplot standardizes X by shifting and scaling each column separately onto the interval [0,1] before making the plot, and centers the glyphs on a rectangular grid that is as close to square as possible. glyphplot treats NaNs in X as missing values, and does not plot the corresponding rows of X.

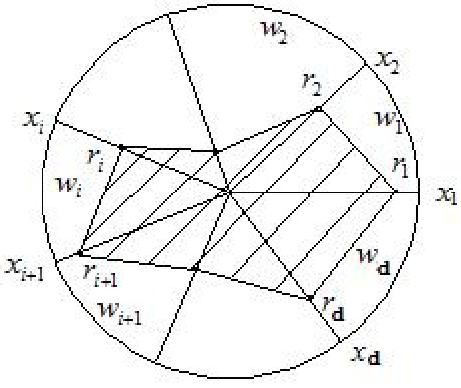

Fig.1 the graphical feature of star plot

B. Graphical Features

When there are many variables involved in a star plot, there is a serious question as to whether a viewer can get a visual impression of the behaviour of a particular variable, or of the joint behaviour of two variables. One of the main purposes of such a scheme is to obtain a star with a distinctive shape for each observation, so that the viewer can look for pairs or groups of stars with similar shapes, or individual observations that are very different from the rest.

The barycentre graphical features are considered as the following. For one observation with n dimension variance, its star plot include n triangle, which is a visionally shape feature. Each triangle has an barycentre G_i,=(abs_i , $angle_i$), and a whole star plot has n barycentre with n amplitude value abs_i and n angle value $angle_i$. So the barycentre graphical features with n amplitude value and n angle value can calculated as the following equation

$$\begin{cases} abs_i = \sqrt{(\frac{r_i}{3}\sin\omega_i)^2 + ((r_i\cos\omega_i - \frac{r_{i+1}}{2})/3 + \frac{r_{i+1}}{2})^2} \\ \qquad angle_i = ar\sin(\frac{\frac{r_i}{3}\sin\omega_i}{abs}) + 2\pi(i-1)/d \\ i = 1, \cdots, d \end{cases} , \qquad (2)$$

For simplification or dimension reduction, we only consider the n amplitude value as the barycentre graphical features for a star plot. Finally the original data are changed to the barycentre graphical features with the same size.

III. FEATURE ORDER BASED ON GENETIC ALGORITHMS

The graphic feature of star plot is affected by the feature order. From the experimental results, feature order not only has affect on the graphic feature, but also has effect on the performance of the classifiers commonly used. To choose d features from features d and require different feature permutation order, all of possible combinatorial numbers is q ＝d!. So for high dimension data, the calculation amount for searching a optimal feature order is too large to realize.

Genetic Algorithms (GA) is randomized search and optimization techniques guided by the principles of evolution and natural genetics. They are efficient, adaptive and robust search processes, producing near optimal solutions and have a large amount of implicit parallelism. The utility of GA in solving problems that are large, multi-modal and highly complex has been demonstrated in several areas. Good solutions are selected and manipulated to achieve new and possibly better solutions. The manipulation is done by the genetic operators (selection, mutation, and crossover) that work on the chromosomes in which the parameters of possible solutions are encoded. In each generation of the GA, the new solutions replace the solutions in the population that are selected for deletion.

We consider integral coded GA for the networks parameters. Each individual is composed of the integer from 1 to the maximum dimension d.

The genetic operation includes selection operation, crossover operation and mutation operation. The selection operation mimics the ‘survival of the fittest’ concept of natural genetic systems. Here the individuals are selected from a population to create a mating pool. The probability of selection of a particular individual is directly or inversely proportional to the fitness value depending on whether the problem is that of maximization or minimization. In this work, the individual with a larger fitness value, i.e. better solution to the problem, receive correspondingly larger numbers of copies in the mating pool. The size of the mating pool is taken to be same as that of population. The crossover operation is a probabilistic process that exchanges information between two parent individual and generates two offspring for the next population. Here one-point crossover with a fixed crossover probability of pc=1 is used. The mutation operation used the uniform mutation. Each individual undergoes uniform mutation with a fixed probability pm =0.005. Elitism is an effective means of saving early solutions by ensuring the survival the fittest individual in each generation. The elitism puts the best individual of old generation into the new generation. A fitness function value is computed for each individual in the population, and the objective is to find the individual that has the highest fitness for the problem considered. The fitness function is the classification correct rate of the classifiers.

As the stopping rule of maximum generation or the minimum criterion of relative change of the fitness values is satisfied, the GA process stops and the solution with the highest fitness value is regarded as the best feature order.

IV. Results and Analysis

A. Data sets Description

For example, in 1991 the Wine dataset was donated into the UCI repository [2]. The data contain 178 examples with measurements of 13 chemical constituents (e.g. alcohol, Mg) and the goal is to classify three cultivars from Italy. This dataset is very easy to discriminate and has been mainly used as a benchmark for pattern recognition.

We also use the data from the UCI repository in 2009[11]. During the preprocessing stage, the database was transformed in order to include a distinct wine sample (with all tests) per row. To avoid discarding examples, only the most common physicochemical tests were selected. Since the red and white tastes are quite different, the analysis will be performed separately, thus two datasets 1 were built with 1599 red and 4898 white examples with 11 dimensions. Regarding the preferences, each sample was evaluated by a minimum of three sensory assessors (using blind tastes), which graded the wine in a scale that ranges from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations.

B. Classification Performance

The confusion matrix is often used for classification analysis, where a C*C matrix (C is the number of classes) is created by matching the predicted values (in columns) with the desired classes (in rows). For an ordered output, the predicted class is given by $p_i = y_i$, if $| y_i – y'_i | <= T$, else $p_i = y_{0i}$, where y_{0i} denotes the closest class to y'_i, given that $y_{0i} \sim= y_i$. From the matrix, several metrics can be used to access the overall classification performance, such as the accuracy and precision. The holdout validation is commonly used to estimate the generalization capability of a model. A more robust estimation procedure is the k-fold cross-validation, where the data is divided into k partitions of equal size. One subset is tested each time and the remaining data are used for training the model. The process is repeated sequentially until all subsets have been tested. Therefore, all data are used for training and testing. However, this method requires around k times more computation.

C. Experiment results

Fig. 1 plots the parallel coordinate plot of the 11 target variables for red wine. Fig. 1 plots the principal component scatter plot of red wine. The plot of the white wine is the same as the red wine. From the figure, the classification task is very difficult.

Table 1 gives the classification results of the red wine and the white wine in 5-fold cross-validation. Our methods works better than the traditional neural networks and support vector machine method. Such model is useful to support the oenologist wine tasting evaluations and improve wine production. Furthermore, similar techniques can help in target marketing by modeling consumer tastes from niche markets.

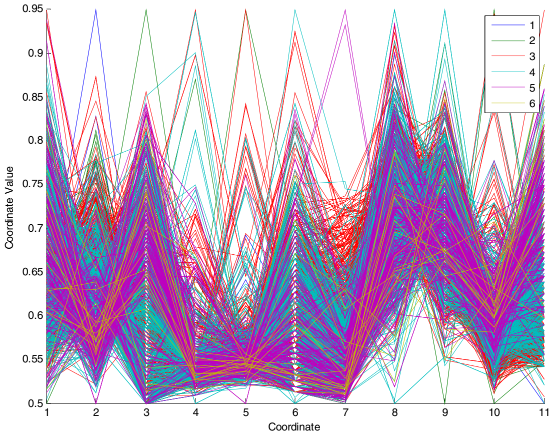

Fig.1 The visual diagram of parallel coordinate plot of red wine

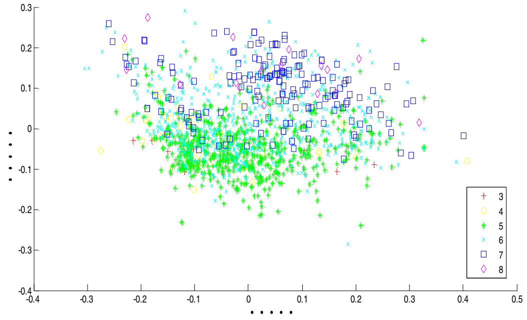

Fig.2 The visual diagram of principal component scatter plot of red wine

Table.1 Table of evaluation results of wine quality

Error (var)	NN	SVM	Graphica l feature	GA-based feature
Red wine	60.0(0.3)	64.3(0.2)	69.1(0.2)	88.4(0.2)
White wine	53.6(0.2)	60.7(0.3)	65.6(0.2)	86.8(0.2)

V. Conclusion

In recent years, interest in wine has increased, leading to the growth of the wine industry. Therefore, companies are investing in new technology to improve production and marketing of wine. Quality is the key factor in play and now mainly rely on human experts in wine. Our work aims to predict from objective analysis of the test preferred the taste of wine. Experiment with encouraging results. Our methods works with classification results 69.1% and 65.6% better than the traditional neural networks and support vector machine method.

The current quality assessment method is based on expert experience and knowledge, are somewhat subjective. The proposed method is based on the objective data-driven model considered integration into a decision support system. Only experts predicted a major departure from the system in the quality evaluation, He can once again taste evaluation.

Once the excavated the relationship between some of the variables and quality evaluation, wine production stage will could control some variables to make the taste better.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No.60504035, 60904100), and the Natural Science Foundation of Hebei Province (No.F2010001281, No.A2010001124). The work was also partly supported by the Doctor Foundation of Yanshan University (B433).

References

[1] http://www.winechina.com/

[2] http://www.ics.uci.edu/~mlearn/MLRepository.html

[3] D. Smith and R. Margolskee. Making sense of taste. Scientic American, 2006,16(3):84-92

[4] Duda R. O., Hart P. E., and Stork D. G. Pattern Classification 2nd ed, vol.3, Wiley, 2000, pp.75-80

[5] Jain, A.K., Duin, R.P.W., Mao, J. Statistical Pattern Recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000,22(1):4–37

[6] Gao, H.X. Application Statistical Multianalysis. Beijing University Press, Beijing, 2005

[7] Jinjia Wang, Wenxue Hong, Xin Li. The new graphical features of star plot for K nearest neighbor classifier. Lecture Notes in Computer Science, 2007(4682):926-933

[8] 19. Jinjia Wang, Jing Li, Wenxue Hong. Feature extraction and classification of Graphical representations of data. Lecture Notes in Computer Science, 2008(5226):534-541

[9] 9. Jinjia Wang, Jialin Song, Xin Li, Wenxue Hong. An Efficient Chernoff Faces Clustering Algorithms. Dynamics of Continuous Discrete and Impulsive Systems, 2006, Vol.13, 1050-1052.

[10] 12. Jinjia Wang, Wenxue Hong, Xin Li. Chernoff Faces Classification Algorithms. Dynamics of Continuous, Discrete and Impulsive Systems, 2007, Vol. 23, pp. 2059-2062

[11] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties[J]. Decision Support Systems, 2009, 47(4):547-553.