Incremental support vector machine combined with ultraviolet-visible spectroscopy for rapid discriminant analysis of red wine

Jun Liu[1], Tie-Jun Pan[2], Zheng-Yong Zhang[1*]

1. School of Management Science and Engineering, Nanjing University of Finance and Economics, Nanjing, Jiangsu, 210023, P. R. China.
2. Ningbo DaHongYing University, Ningbo, Zhejiang, 315175, P. R. China.

[*]Correspondence to Zheng-Yong Zhang, School of Management Science and Engineering, Nanjing University of Finance and Economics, Nanjing, Jiangsu, 210023, P. R. China. E-mail: zyzhang@nufe.edu.cn

Abstract:

The aim of this work is to develop a new method for overcome the increased training time when a recognition model is updated based on the condition of new features are extracted from new samples. As a common complex system, red wine has a rich chemical composition and is used as an object of this research. The novel method based on incremental learning support vector machine (I-SVM) combined with ultraviolet–visible (UV-Vis) spectroscopy were applied to discriminant analysis of the brands of red wine for the first time. In this method, new features included in the new training samples were introduced into the recognition model through iterative learning in each iteration, and the recognition model was rapidly updated without significantly increasing the training time. Experimental results show that the recognition model established by this method obtains a good balance between training efficiency and recognition accuracy.

Key words: Incremental SVM, Identification, Ultraviolet–visible spectroscopy, Red wine

1． Introduction

Ultraviolet–visible (UV-Vis) absorption spectra based on intramolecular electron transitions obtained by UV-scanning a substance over a range of wavelengths with a UV spectrophotometer. The difference of peak shape, peak height and peak area of the UV-visible spectrum characterizes the disparity of the composition and the degree of unsaturation of the components contained in the sample, which reflects the overall characteristics of the sample. UV-Vis spectroscopy has the characteristics of high sensitivity, good reproducibility, high efficiency and low cost. Detection with pattern recognition algorithm has been widely used in food, medicine and other fields, and obtains a good accuracy[1-4].

The traditional method of pattern recognition is a kind of off-line training method, which trains a classifier with labeled sample datasets for recognition. The quality of red wine is mainly determined by the raw materials of the grape and the brewing process, while the quality of the grape raw material is greatly influenced by the climate of the place of production, that makes the tastes of different batches red wines of the same brand have subtle differences, and the corresponding spectral data also change, the classification accuracy of the classifier trained by off-line data will be significantly reduced. The solution is usually to retrain the classifier, but retraining requires a large number of training samples and training time, this cost is unbearable. Therefore, how to adapt the classifier trained by the old labeled data to the identification of new samples is a difficult problem in on-line identification [5-8].

Support Vector Machine (SVM) has been successfully used for data mining, pattern recognition and artificial intelligence fields. With labeled data, SVM learns a boundary (i.e., hyperplane) separating different class data with maximum margin. The classification process usually face the new evolving data, the initial training sample data cannot reflect all the sample information. When new training samples are accumulated to a certain scale, in order to obtain the new sample information, it would like to integrate these examples and train a new classification model. However, the training of a SVM has the time complexity of $O(n^3)$(n is the number of training samples), it does benefit large-scale online applications [9-12].

It is noteworthy that performance of classification method for red wine is evaluated not only based on accuracy, but also the rapidity, which are also of great significance in practical applications. To attack this problem, lots of works have been done. One way is to reduce training samples with a certain sample selection strategy. The quality of training data set is vital to the performance of the classifier being constructed. Syed et al. worked out an incremental algorithm based on SVM, which retains only the support vector set as a historical training sample.

The main contribution of this paper is that a novel hybrid classification method based on Principal Components Analysis (PCA) and Incremental Support Vector Machine (I-SVM)[12-13], combined with UV-Vis spectra is proposed. Experimental results indicated that PCA-I-SVM, as a classifier, was tested in terms of classification rate and running time. Compared with normal SVM, PCA-I-SVM can run much faster with similar accuracy rate. Experimental results showed that PCA-I-SVM combined with UV-Vis spectra can be a rapid, accurate method for classification of red wine.

2. Experiments and Materials

2.1 Sample collection and preparation

Nine brands of red wine, a total of 54 wine samples, were purchased from a well-known e-commerce website in China. At the same time, in order to verify the time efficiency of incremental learning algorithm, a total of 5400 samples of 100 batches of samples were simulated by Monte Carlo method.

2.2 UV-Vis spectrum acquisition

The ultraviolet-visible (UV-Vis) spectra were obtained from a UV-Vis spectrometer, T6 New century, Purkinje General Instrument Co., Ltd. (Beijing, China). Water was used as the zero point. Each test sample was prepared through mixed 100 μL wine with 3 mL water. The scanning range of the UV-Vis spectrum of each sample was 240~550 nm.

2.3 Data preprocessing

In this study, we used PCA to remove redundant features and several previous principal components were extracted as the inputs of the classifier for red wine. PCA is a method for the re-expressing multivariate data. It allows the researcher to reorient the data so that the first few dimensions account for as much of the available information as possible. The principal components solution has the property that each component is uncorrelated with all others, which has the advantage of eliminating multicollinearity.

The number of the generated features was still quite large for the classifier. So PCA was used to perform feature reduction before pattern recognition, then I-SVM was used for classification of red wine.

2.4 Incremental SVM Learning based on Support vector

In order to make the SVM learning algorithm incremental, reducing the number of training data sets is an effective way to apply SVM classification for large data sets. Because Support Vectors are a sufficient description of the decision boundary between the examples, then at each incremental step, the representation of the old sample data is given by the set of Support Vectors. Such support vectors are incorporated with the new incoming batch of data to provide the training data for the next step. Since the number of support vectors to be small compared to the total number of training examples, this method can effectively reduce the number of samples.

Now suppose that labeled data $\mathrm{X} = \{x_{ij} | i = 1,2,\dots,c, j = 1,2,\dots,N_i\}$ is the UV-Vis spectra data of , where c is the red wine class number, x_{ij} represents the jth samples in class i, N_i is the number of samples in class i . The overall red wine sample size is N, which is expressed by:$\mathrm{N} = \sum_{i=0}^{c} N_i$.

The incremental SVM algorithm framework as fellows, see Figure 1:

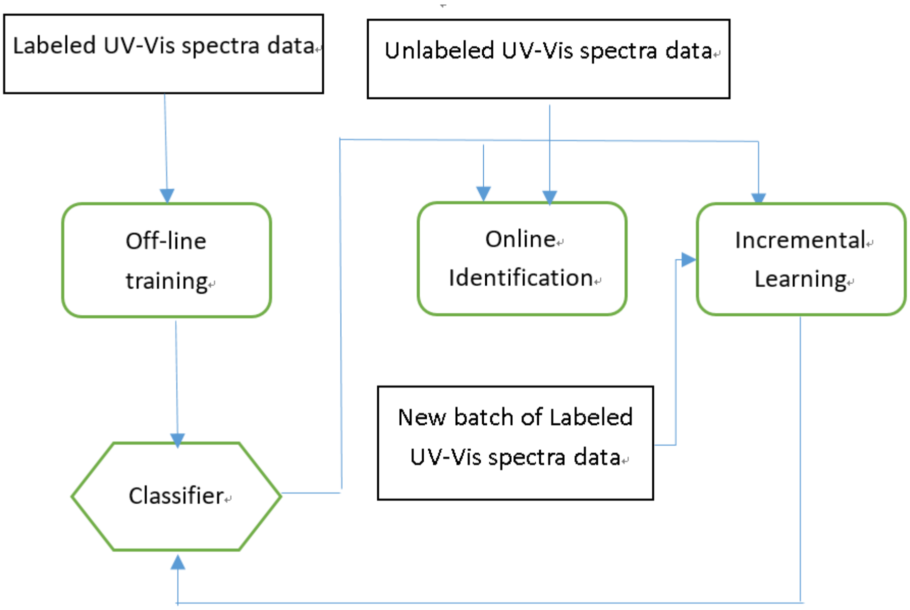
Labeled UV-Vis spectra data
Unlabeled UV-Vis spectra data
Off-line training
Online Identification
Incremental Learning
New batch of Labeled UV-Vis spectra data
Classifier

Figure 1 Incremental Learning SVM Algorithm Framework

1. Uses labeled UV-Vis spectra data X_{ini} to initialize SVM classifier.

2. For each increment learning step. Creating subsets $X^i\{,i = 1,2,, \dots, c\}$ from UV-Vis spectra data set X. for each class data set X^i, compute the Support Vector set X_{SV}^i, then concatenate $X_{SV} = X_{SV}^1 \cup X_{SV}^2 \cup \cdots \cup X_{SV}^c$.

3. Uses X_{SV} and new labeled UV-Vis spectra data as train data set to train SVM model, and get Support Vectors set X_{SV}, then get the updated classifier.

4. Repeat steps 3) and 4) enable continuous incremental learning of new batch labeled UV-Vis spectra data.

2.5 Model validation

To assess the performance of the established classifier, leave-one-out cross-validation and 10-fold cross-validation were conducted. These cross-validations fully assessed the performance of the classification model.

3. Data Analysis

UV-Vis spectra can quickly obtain sample information about the functional groups in aromatic compounds, and has significant advantages that include simple sample preparation, rapid analysis, high sensitivity, robustness, green process, and low cost.

Since the UV-Vis spectra mainly reflected the main compounds of the red wine, UV-Vis spectra of red wine are extremely similar and difficult to be identified manually. As shown in Figure 2. The spectrum of red wine at 240~550 nm are similar. The UV-Vis absorption around 260~275 nm could be attributed to pi-pi* electron transition between bonding pi orbital to antibonding pi* orbital, for example unsaturated hydrocarbon and aromatic hydrocarbons. In traditional spectral analysis, Beer-Lambert law was used commonly, that is, the maximum

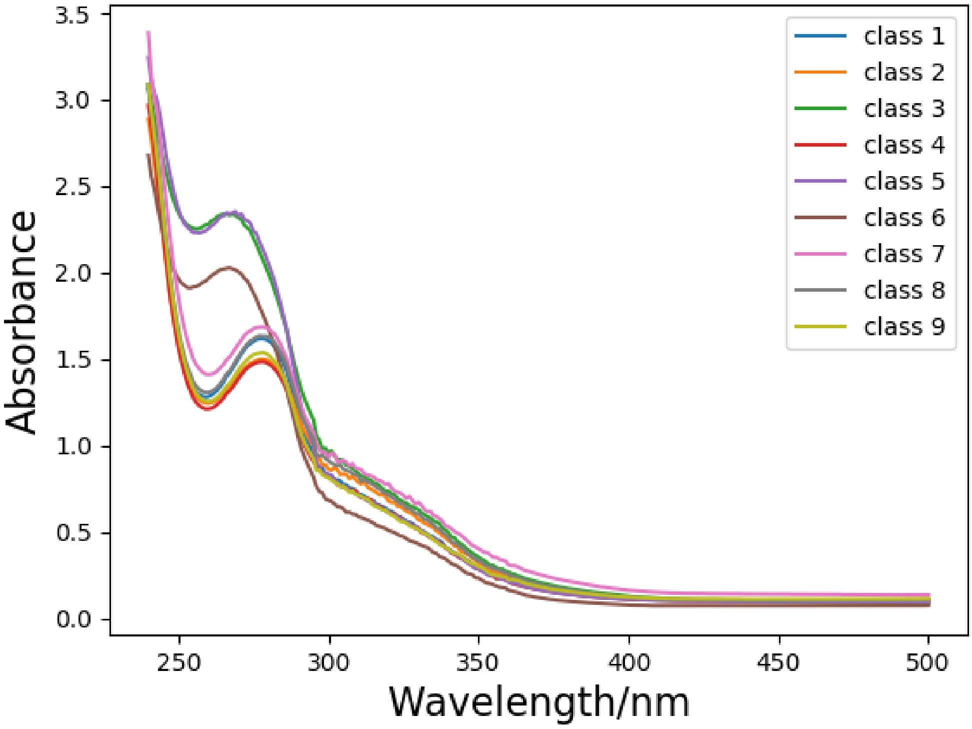

Figure 2: Typical UV-Vis spectra of nine kinds of red wine

absorption peak and the concern between specific substances were applied to analyze, and the utilization rate of spectral information was relatively low. In fact, the information of the UV-Vis spectrum of red wine is very rich, its peak shape, area, width and so on are closely related to the quality of the sample. However, if we only rely on the observation of the spectrum, it is not appropriate to make a visual judgment of the quality of the red wine, because most of the time different molecules contribute to the similar peak. In order to overcome this limitation of visual analysis, statistical methods are mostly used for the further analysis of the UV-Vis data. With the statistical approach one can extract useful information from the data set by high lighting the similarities and differences. In this study, we used I-SVM for the classification of red wine, in order to efficiently handle large amounts of sample data.

4 Result and Discussion

4.1 Raw data of Characteristic Information

The chemical components and relative contents of different flavors are different, these will produce different associations, so it determine the spectral curves of different flavors are somewhat different, and has different characteristics and fingerprints. The difference between the spectra is the variation of relative intensities of the absorption peaks in the fingerprint region, and the minute difference in the small peaks in the fingerprint region. Pattern recognition algorithm can maximize the information extracted from the data, and can classify the sample set.

4.2 Red wine classification

4.2.1 Data preprocessing results with PCA

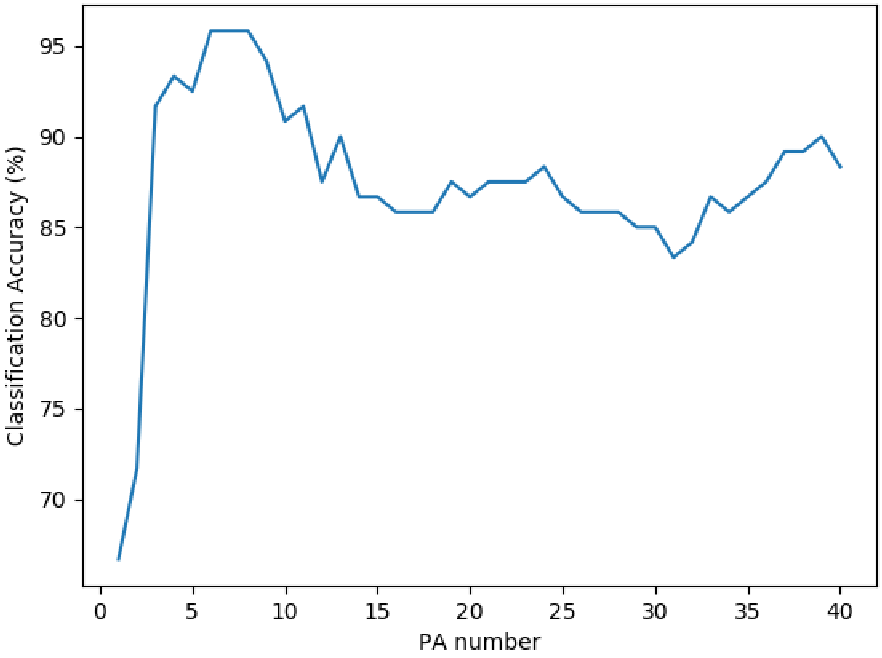

Figure 3: The Classification rate of principal component number

The dimension of feature space generated by PCA is not determined by itself, and depended on the final classification rate and efficiency, according to Figure 3, we utilize 8 principal components as feature vectors, thinking of account the balance between efficiency and classification accuracy.

4.2.2 Classification result with I-SVM

The I-SVM algorithm was used to classify the nine brands of red wine samples. We selected the RBF kernel function in the I-SVM algorithm, and the kernel parameter was optimized using grid search with cross-validation method. We used leave-one-out cross-validation and 10-fold cross-validation to assess the performance of these classifier.

4.2.3 Compare data processing efficiency and accuracy

For comparison, three different algorithms were simulated. Algorithm 1 uses the normal SVM algorithm, which uses all the samples to solve the support vector for each incremental learning. Algorithm 2 is I-SVM algorithm, which uses the support vector set for incremental learning. Algorithm 3 is the Multi-layer Perceptron Neural Network (MLPNN) algorithm. The initial sample set is 300 samples randomly selected from all samples, and 510 samples are added for each incremental learning. The results are shown in Figure 4.

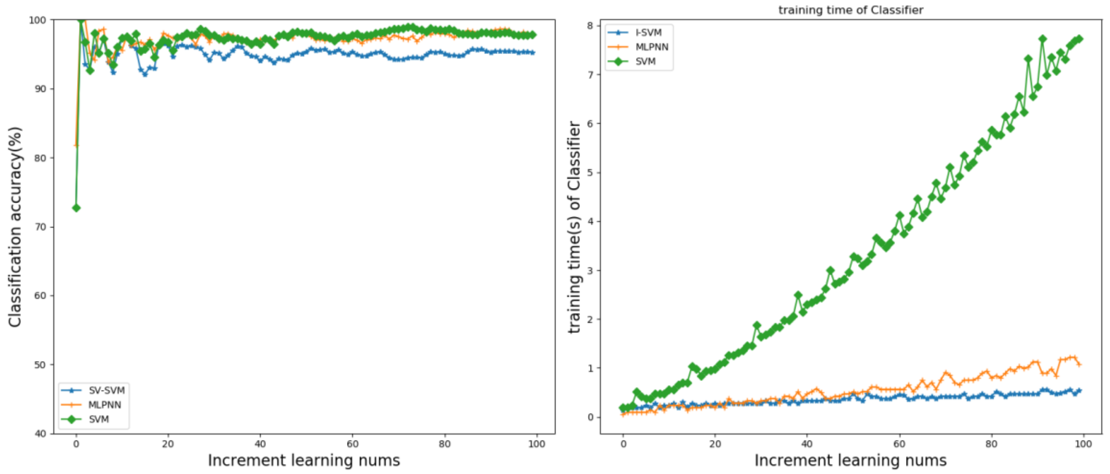

Figure 4: Comparison of three algorithms

Support Vector set with the new sample set rather than an initial sample set, greatly saves the computation time and accelerates the simulation speed, and the classification accuracy is basically the same. Meanwhile, with the continuous learning of incremental learning, the algorithm can naturally make part of the Support vectors into non-SV vector, to achieve the selective forgetting of the historical data of the training. Therefore, when dealing with a large number of new training data, the speed advantage of the incremental I-SVM method is more remarkable.

5. Conclusion

The experiment will be UV-visible spectroscopy and I-SVM combination of red wine used online identification. After the pretreatment of the samples of red wine, the ultraviolet spectral fingerprint library of nine kinds of red wine was established by UV spectrophotometer. After dimensionality reduction by PCA, an incremental SVM model was established to identify the red wine. The recognition rate of red wine reached 94.9%. At the same time, in order to verify the time efficiency of the algorithm, a total of 5400 samples of 100 batches of samples were simulated by Monte Carlo method. The recognition rate of wine reached 96.78%. The average training time of I-SVM models for each batch was 0.47seconds with standard deviation of 0.03, one-thirteenth of the average time of normal SVM. The method provides a reliable, stable, rapid and completely new method for the identification of online red wine, and provides a method basis for quality evaluation and quality control of red wine.

6. Acknowledgements

This paper was supported by the National Natural Science Foundation of China (No.61373058, 61602217).

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] SHA Min, SONG Chao, ZHANG Zhengyong et al. Discrimination of Red wines Based on Spectral Data Fusion Combined with Pattern Recognition Algorithm. FOOD SCIENCES, 2016,37(22):192-197

[2] M.J.Martelo-VidalM.Vázquez, Determination of polyphenolic compounds of red wines by UV–VIS–NIR spectroscopy and chemometrics tools, Food Chemistry, **2014,**158 :28-34

[3] María J. Martelo-Vidal, Manuel Vázquez,Classification of red wines from controlled designation of origin by ultraviolet-visible and near-infrared spectral analysis, Ciência Téc. Vitiv. 2014,29(1):35-43.

[4] Alamprese C., Casale M., Sinelli N., Lanteri S., Casiraghi E.,. Detection of minced beef adulteration with turkey meat by UV-vis, NIR and MIR spectroscopy. Lwt-Food Sci. Technol., 2013,53,:225-232.

[5] Luna A.S., da Silva A.P., Ferre J., Boque R., Classification of edible oils and modeling of their physico-chemical properties by chemometric methods using mid-IR spectroscopy. Spectrochim. Acta A, 2013,100:109-114.

[6] Abdi H., Williams L.J. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat.,2010,2:433-459.

[7] Zhou Xiu jun,Dai Lian kui,Li Sheng. Fast Discrimination of Edible Vegetable Oil Based on Raman Spectroscopy. Spectroscopy and Spectral Analysis Spectrosc Spect Anal,2012,32(7):1829-1833

[8] K.P. Bennett , E.J. Bredensteiner, Geometry in Learning, Geometry at Work, C. Gorinieditors, Mathematical Association of America, 2000,132-145

[9] Y.J. Lee, S.Y. Huang, Reduced support vector machines: a statistical theory, IEEE Transactions on Neural Networks. 2007, 18 (1):1–13.8

[10] Pavel Laskov, Christian Gehl, Stefan Kruger. Incremental Support Vector Learning: Analysis, Implementation and Applications. Journal of Machine Learning Research. 2006,7:1909–1936

[11] GU B，ZHeng G S，Wang J D. Analysis for Incremental and Decremental Standard Support Vector Machine． Journal of Soft-ware. 2013,4(7) : 1601-1613．

[12] Tsang, J. Kwok, and P.-M. Cheung. Core Vector Machines: fast SVM training on very large data sets. Journal of Machine Learning Research. 2005,6:363–392

[13] M. J. Tax and P. Laskov. Online SVM learning: from classification to data description and back.In C. et al. Molina, editor, Proc. NNSP,.2003,499–508