# Rapid Discrimination of apple essence base on PCA-CH-SVM

刘军 (NUFE,210042)

Abstract:An important aspect of rapid discrimination of apple essences based on pattern recognition is how to use new training data to improve the accuracy and control the training time. In this paper,In this paper, PCA-CH-SVM method and Raman spectra were used in combination for fast discrimination of apple essences from different brands.PCA-CH-SVM built on a convex hull of support vectors and new Raman spectra data to classify the apple essence based on features obtained from Raman spectra.The classification model have been evaluated with 10-fold cross validation.The results from this study demonstrated that our approach has good classification accuracy while the training is significantly faster than normal SVM classifiers.

In this paper,we proposed a incremental learning algorithm by combining convex-hull SVM with PCA,namely PCA-CH-SVM for incremental learning in apple essences discrimination.

Key words:Apple Essences;Convex hull vector; SVM ; rapid discrimination

## 1 I. Introduction

Apple essences are widely used as a food additive in food industry.Rapid identification of apple essence for the food industry's quality control is of great significance.Apple essences are a complex mixture of a large number of volatile compounds[1].Usually the detection of apple essence is carried out by chemistry-based methods and sensory evaluation.sensory evaluation is the traditional and most commonly used method,but its accuracy and objectivity cannot always be ensured because sensory evaluation staff' s judgement can affected by their health condition, emotions, and the environment.

These chemistry-based methods such as gas chromatography, mass spectrometry, and gas chromatography-mass spectrometry [5–8] are highly reliable because they use a complete component-by-component approach.However, their shortcomings include excessive test items, being time-consuming,complicated operation, and low capability for insitu and rapid measurements[9] . Overall, developing a novel, rapid and reliable method to identify essence is of positive significance.

Raman spectroscopy is a technique which is arising from inelastic scattering of laser light by the molecular vibration inside the sample. As a result, the scattered photons are emitted with the different frequency or energy. This difference in frequency between incident and emitted protons provides finger print about the rotational, vibrational and other low frequency transitions in molecule. Thus Raman spectrum, which is the plot of intensity as function of Raman shift, is a rapid detection method developed in recent years, with fast, efficient, non-polluting, without pre-treatment, lossless analysis, etc., many areas have been widely used.[拉曼在食品检测中的文献]

Support Vector Machine (SVM) [2] has been successfully used for for data mining, pattern recognition and artificial intelligence fields [2–5]. 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 set can not 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 $O(M^3)$(M is the number of training samples), it does benefit large-scale online applications.

It is noteworthy that performance of classification method for apple essence is evaluated not only recognition accuracy,but also the rapidity,which is 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.[3] 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,PCA-CH-SVM,combined with Raman spectra is proposed.Experimental results indicated that PCA-CH-SVM ,as a classifier,was tested in terms of classification rate and running time.Compared with normal SVM,PCA-CH-SVM can run much faster with similar accuracy rate.Experimental results showed that PCA-CH-SVM combined with Raman spectra can be a rapid,accurate method for classification of apple essences.

# 2 II.Experiments and Materials

## 2.1 Sample collection and preparation

A total of 27 experimental samples, corresponding to 3 apple essence brands,were obtained from three famous flavors and fragrances companies in China by three batches.All samples were produced in 2016,and had equivalent proofs.The apple essence included in study are listed in Table1.

The overall procedure of sample collection is same. In total,Raman spectra of 300 samples of 3 apple essence brands have been used in this study.Out of 300 samples,xx were

Essence contains a large number of volatile, low content components.The complex pretreatment methods of samples have some impact on these components.In order to avoid introducing other impurities or the distortion of component proportion caused by improper pretreatment method,in this experiment, the test samples are prepared by high dilution of pure water.Add 3 grams of essence in the volumetric flask,was respectively diluted 10 times and 1000 times with high purity water,and shaked well, then got samples.The standard safety rules have been followed at each step from sample collection till acquisition of Raman spectra。

## 2.2 Raman spectrum acquisition

Raman spectrum for all samples have been acquired with Raman spectrometer(ProttezRaman-d3,Enwave Optronics, USA).Raman signal is normally very weak as compared to Rayleigh scattering, therefore an acquisition time of 10 seconds has been used for recording each spectrum.The spectrum from the sera samples have been recorded in the spectral range of 250 $cm^{-1}$ to 2350 $cm^{-1}$, as it contained the most useful information.

Table 1: Detailed information of the investigated apple essences

| flavor companies | no. | solvent |
|---|---|---|
| A | S | ethanol |
| | Q | ethanol |
| | I | 1,2 propanediol |
| B | a | 1,2 propanediol |
| | b | ethanol,1,2 propanediol |
| | c | 1,2 propanediol,water |
| C | d | 1,2 propanediol |
| | e | ethanol |
| | f | 1,2 propanediol |

## 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 apple essences. Principal components analysis(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 standard soft-margin SVM was used for classification of apple essences.

## 2.4 Incremental SVM Learning Base on Convex Hull Vector

Reducing training data sets is an effective way to apply SVM classification for large data sets.The geometric properties of SVM can also be used to reduce the training data.The maximum-margin hyperplane is written in terms of data instances that belongs to the outside of the boundaries of the classes.in the separable case,the boundaries of classes contain the instances of solution(support vectors),therefore we only need the points on those boundaries.The boundaries of the data can be obtained from the Convex Hull of each class.

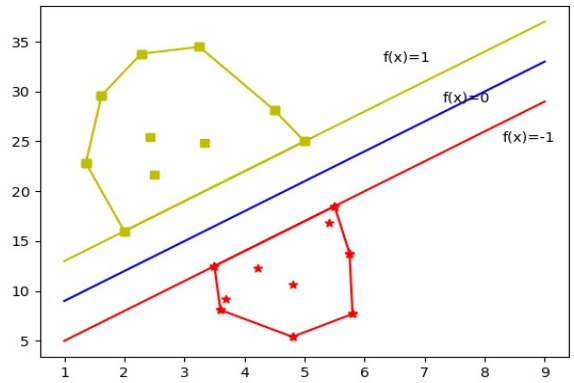


Figure 1: Relationship between hull vectors and support vectors

Asdrúbal López Chau et al. [4] proposed convex-concave hull SVM classifier has distinctive advantages on dealing with large data sets with higher accuracy.As described in this paper,the vertices of the convex-concave hull are applied for SVM training with higher accuracy.In this work,a Convex Hull SVM algorithm was used for classification due to its good incremental learning.

Now suppose that $X = \{x_{ij}|i = 1, 2, ..., c, j = 1, 2, ..., N_i\}$ is the Raman spect data of apple essences,where c is the apple essences class number,$x_{ij}$ represents the $j$th samples in class $i$,$N_i$ is the number of samples in class i.The overall apple essences sample size is N,which is expressed by

$$N = \sum_{i=1}^{c} N_i$$

The convex hull(CH) of a set of points S is the minimum convex set that contains S.Mathematically,CH is defined as:

$$CH(X) : \{\omega = \sum_{i=1}^{n} \alpha_i x_i, \alpha_i \geq 0, \sum_{i=1}^{n} = 1, x_i \in X\}$$

Firstly, Let's create two subsets $X^+$ and $X^-$ from $X$ The procedure is summarized as follows:

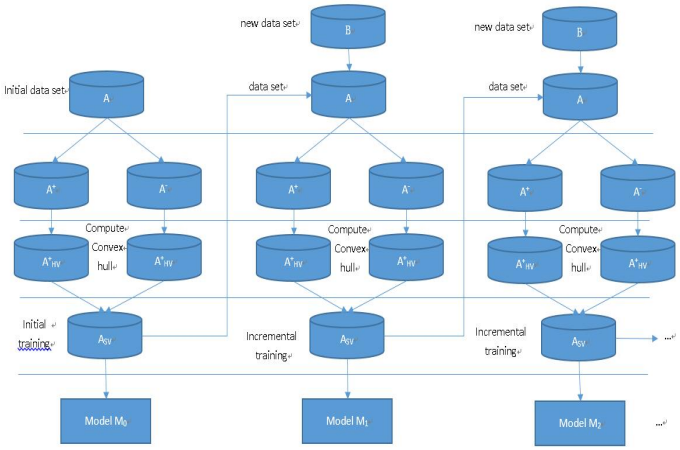


Figure 2: Gneral process of the method,take two classes as sample

1. Firstly,create subsets $X^i, i = 1, 2, ..., c$ from $X$.for each class data set $X^i$ ,compute the convex hull vector set $X^i_{H_V} = CH(X_i)$,then make $X_{H_V} = X^1_{H_V} \bigcup X^2_{H_V} \bigcup ... \bigcup X^c_{H_V}$

2. Use $X_{H_V}$ as train data set, train SVM model,and get support vectors $X_{S_V}$

3. Add the incremental train data set S,splits into subsets $S^i, (i = 1, 2, ..., c)$ from $S$. make $X^i = S^i \bigcup X^i_{H_V}$ ,compute the convex hull vector set $X^i_{H_V} = CH(X_i)$

4. As hull vector set $X_{H_V}$ as train data set to train SVM model,and get support vector set $X_{S_V}$ ,then get the classifier. Repeat steps 3) and 4) enable continuous incremental learning of new samples.

# 3 III. Data Analysis

## 3.1 Raman spectrum Data analysis

Raman spectroscopy can quickly obtain sample information about the functional groups in aromatic compounds,and has significant advantages: sample preparation is simple, measurement usually does not destroy sample, and moisture does not affect test.

Raman spectrum of Apple essence samples is normally very complex and rich of information of functional group of organic compounds.The Raman spectra mainly reflected the solvent information of the essence, and the Raman spectra of the essence with the same solvent are extremely similar and difficult to be identified manually,as shown in Figure 1.The spectra of essences e, Q, and s are similar,and i,a,c,d,f is similar.The spectra B has more peaks,and contains the peaks of the previous two types of spectra.According to the literature[23] and comparison of standards,the spectra of essences e, Q, and s are the peak of

Raman spectrum of essence samples is normally very complex and rich of chemical information. Since in essence samples, there exist different types of functional group compound. The Raman spectrum of each of these compound consists of numerous peaks. The visual assignment of any particular peaks to a specific molecule usually produces imprecision in the final result, because most of the time different molecules contribute to the same peak. In order to overcome this limitation of visual analysis, statistical methods are mostly used for the interpretation of Raman data of essence samples. With the statistical approach one can extract useful information from the data set by high lighting the similarities and differences. In this study, we are using convex hull SVM for the classification of apple essence, in order to efficiently handle large amounts of sample data.

# 4 IV. Result and Discussion

## 4.1 Raw data of Characteristic Information

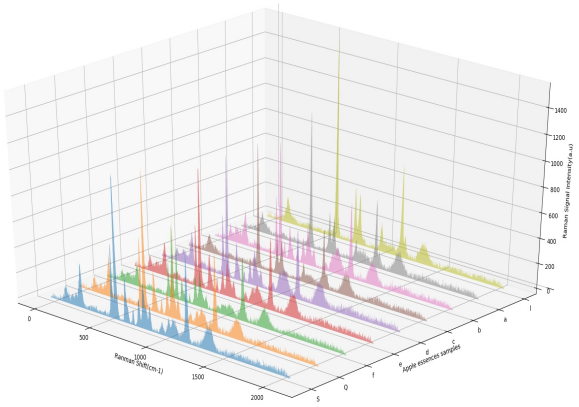


Figure 3: Typical Raman spectra of nine kinds of apple essences

Taking nine kinds of apple essences as example,figure x shows Raman spectra of apple essences.In the range of 1500 2350 $cm^{-1}$ band, the peak is small and mostly the background peak,so only the Raman spectral data in the 350 1500 $cm^{-1}$ band is considered for data

processing.So the peak corresponding to 350 1500 $cm^{-1}$ band constitutes a feature vectors with the size of $1 \times 1150$.

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 Apple essences classification

### 4.2.1 Data preprocessing results with PCA

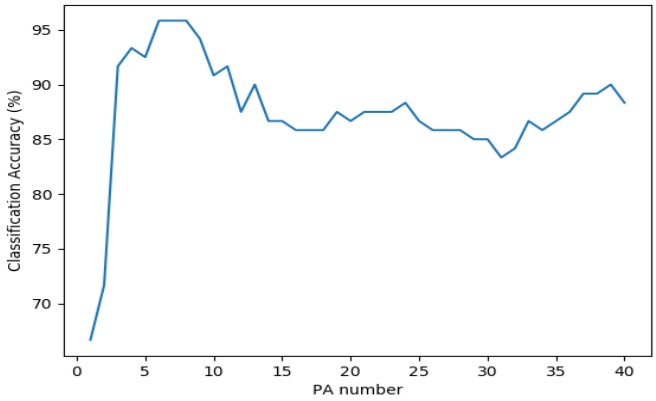


Figure 4: The Classification rate of principal component number

Taking nine kinds of apple essences as example,figure x shows Raman spectra of apple essences.In the range of $1500 \sim 2350cm^{-1}$ band, the peak is small and mostly the background peak,so only the Raman spectral data in the $350 \sim 1500cm^{-1}$ band is considered for data processing.So the peak corresponding to $350 \sim 1500cm^{-1}$ band constitutes a feature vectors with the size of $1 \times 1150$. The dimension of feature space generated by PCA is not determined by itself, and depended on the final classification rate and efficiency,according to Fig 4, We utilize 8 principal components as feature vectors, thinking of account the balance between efficiency and classification accuracy.

### 4.2.2 Classification result with CH-SVM

The CH-SVM algorithm was used to classify the ten brands of apple essences samples. We selected the RBF kernel function in the CH-SVM algorithm, and the kernel parameter was optimized using the Particle Swarm Optimization(PSO) method. 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.

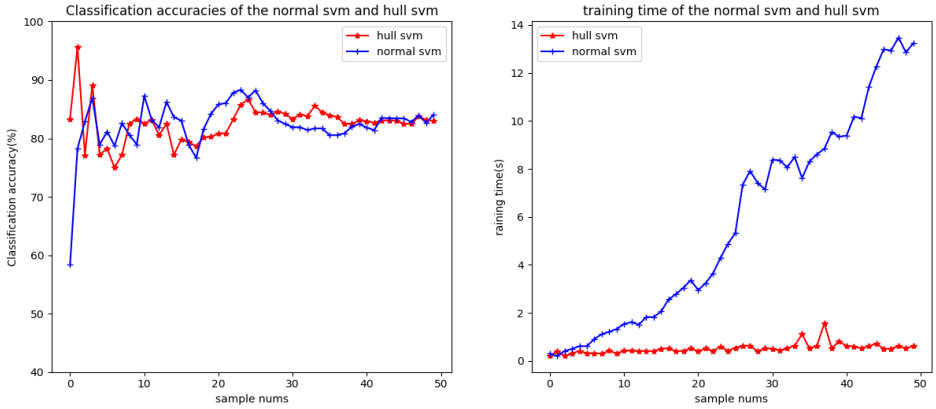


Figure 5: Typical Raman spectra of nine kinds of apple essences

### 4.2.3 Compare data processing efficiency and accuracy

For comparison, three different algorithms were simulated.Algorithm A uses the standard SVM algorithm, which uses all the samples to solve the support vector for each incremental learning.Algorithm B uses the support vector set instead of the original sample set for incremental learning,Algorithm 2 uses the support vector set instead of the original sample set for incremental learning, that is,using the original classifier support vector set instead of the original sample set, combined with the new sample to be calculated together.Algorithm 3 is a shell vector incremental learning algorithm.The initial sample set is 135 samples randomly selected from all samples, and 70 samples are added for each incremental learning. After each learning, using all 345 samples to check the classification effect.The results are shown in Table 2,after the first to third incremental learning process,Compared with the new shell vector set, the number of shell vectors transformed into non-shell vectors is 14,17,18.

Table 2: Simulation results after adding group samples

| learning process | algorithm | Simulation sample | Hv | Sv | t/s | $\eta$ |
|---|---|---|---|---|---|---|
| Initialization(randomly | 1 | 135 | - | 34 | 37.5 | 85.80 |
| selected 135 sample | 2 | 135 | - | 34 | 37.5 | 85.80 |
| data) | 3 | 135 | - | 34 | 37.5 | 85.80 |
| After the first | 1 | 135 | - | 34 | 37.5 | 85.80 |
| incremental study(add | 2 | 135 | - | 34 | 37.5 | 85.80 |
| 70 sample data) | 3 | 135 | - | 34 | 37.5 | 85.80 |
| After the second | 1 | 135 | - | 34 | 37.5 | 85.80 |
| incremental study(add | 2 | 135 | - | 34 | 37.5 | 85.80 |
| 70 sample data) | 3 | 135 | - | 34 | 37.5 | 85.80 |
| After the third | 1 | 135 | - | 34 | 37.5 | 85.80 |
| incremental study(add | 2 | 135 | - | 34 | 37.5 | 85.80 |
| 70 sample data) | 3 | 135 | - | 34 | 37.5 | 85.80 |

It can be seen from the simulation results that the SVM incremental learning algorithm based on shell vector is compared with the standard SVM method, which greatly saves the com-

putation time and accelerates the simulation speed, and the classification accuracy is basically the same,the algorithm, that combined the original 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 Hull vector into non-Hull vector, to achieve the selective forgetting of the historical data of the training.Therefore, when dealing with a large number of online data set ,the speed advantage of the incremental hull SVM method is more obvious.

# 5 V. Conclusion

This study demonstrates the use of Raman spectroscopy combined with Convex-Hull SVM technique for the classification of the spectral data acquired from Apple essense. Raman spectroscopy coupled with statistical tools has great potential to contribute significantly in the On-line inspection and research of product quality in an effective way.There is also a great likelihood to use Raman spectroscopy combined with one of the existing methods for initial screening in order to increase the inspection efficiency. The results obtained are quite promising and interesting. The research work in our laboratory is still in progress striving for increasing sensitivity as well as specificity.

# References

[1] Sanz, C, Olias, J. M, Perez, A. G. Aroma bio-chemistry of fruits and vegetables. In: Tomas-Barberan, F. A.; Robins, R. J. ed. Phytochemistryof fruit and vegetables. New York, Oxford Uni-versity Press Inc.1997,125-155.

[2] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York,1995, 8 (6) :988 - 999

[3] Stefan Ruping,Incremental Learning with Support Vector Machines,Technical Reports,2001,228(4):641-642

[4] Asdrúbal López Chau, Xiaoou Li, Wen Yu, Convex andconcave hulls for classification with support vector machine, Neurocomputing,http://dx.doi.org/10.1016/j.neucom.2013.05.040

[5] XIAO Rong, WANG Ji-cheng, SUN Zheng-xing. Anapproach to incremental SVM learning algorithm.Journal of Nanjing University, 2002, 38(2): 152 157.

[6] Zhu X, Lafferty J, Ghahramani Z. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions[C]. In: Proc of ICML 2003 Workshop on the Continuum from Labeled to Unlabeled Data. Menlo Park, CA:AAAI Press,2003:58-65.

[7] Hyunjung Shin, Sungzoon Cho, Invariance of neighborhood relation underinput space to feature space mapping, Pattern Recognition Letters, 26 (2005)707–718.

[8] Y.J. Lee, S.Y. Huang, Reduced support vector machines: a statistical theory,IEEE Transactions on Neural Networks 18 (No.1) (2007) 1–13.

[9] Jing, Y.; Meng, Q.; Qi, P.; Zeng, M.; Li, W.; Ma, S. Electronic nose with a new feature reduction method and a multi-linear classifier for Chinese liquor classification. Rev. Sci. Instrum. 2014, 85, 055004.

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

[11] K.P. Bennett , E.J. Bredensteiner, Duality and Geometry in SVM Classifiers, 17thInternational Conference on Machine Learning, San Francisco, CA, 2000