Journal of Pharmaceutical and Biomedical Analysis 48 (2008) 1321–1325

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Identification of green tea's (*Camellia sinensis* (L.)) quality level according to measurement of main catechins and caffeine contents by HPLC and support vector classification pattern recognition

Quansheng Chen*, Zhiming Guo, Jiewen Zhao

School of Food & Biological Engineering, Jiangsu University, Zhenjiang 212013, PR China

ARTICLE INFO

Article history:
Received 25 July 2008
Received in revised form 5 September 2008
Accepted 6 September 2008
Available online 17 September 2008

Keywords:
High performance liquid chromatography (HPLC)
Green tea
Quality level
Identification
Support vector classification (SVC)

ABSTRACT

High performance liquid chromatography (HPLC) was identified green tea's quality level by measurement of catechins and caffeine content. Four grades of roast green teas were attempted in this work. Five main catechins ((−)-epigallocatechin gallate (EGCG), (−)-epigallocatechin (EGC), (−)-epicatechin gallate (ECG), (−)-epicatechin (EC), and (+)-catechin (C)) and caffeine contents were measured simultaneously by HPLC. As a new chemical pattern recognition, support vector classification (SVC) was applied to develop identification model. Some parameters including regularization parameter (R) and kernel parameter (K) were optimized by the cross-validation. The optimal SVC model was achieved with $R = 20$ and $K = 2$. Identification rates were 95% in the training set and 90% in the prediction set, respectively. Finally, compared with other pattern recognition approaches, SVC algorithm shows its excellent performance in identification results. Overall results show that it is feasible to identify green tea's quality level according to measurement of main catechins and caffeine contents by HPLC and SVC pattern recognition.

1. Introduction

Green tea (*Camellia sinensis* (L.)) is one of the most popular beverages across the world, which is of great interest due to its beneficial medicinal properties [1]. There is increasing evidence that specific substances found in the tea leaves can enhance general health. Recent research suggests that some compounds including caffeine and catechins found in tea leaves may play an important role to prevent cardiovascular disease [2], chronic gastritis [3,4] and some cancers [5,6]. There are many catechins in green tea leaves, and principal catechins existing in green tea leaves are (−)-epigallocatechin gallate (EGCG), (−)-epigallocatechin (EGC), (−)-epicatechin gallate (ECG), (−)-epicatechin (EC), and (+)-catechin (C). The chemical structures of five catechins are shown in Fig. 1. As shown in their chemical structures, these catechins exhibit high antioxidation level. With the increasing consumption of tea, quality control of tea becomes more and more important nowadays, for example, many national and international authorities are setting criteria for quality index. The most important chemical constituents that influence the taste and flavor of tea infusions are polyphenols, flavonols, caffeine, sugars, amino acids, and volatile flavor compounds in green tea leaves [7]. It has been demonstrated that there is a relationship between the green tea's quality level and its main catechins contents. The concept of catechins index (CI) was offered to estimate the quality of green tea. These compounds are mainly responsible for the characteristic astringent and bitter taste of tea brews. Besides these catechins, caffeine is another important factor in determining the quality level of green tea. Caffeine is the main methyl xanthine constituting the tea alkaloids. It is recognized as an important quality factor because of its stimulative effect. In contrast to the catechins in green tea leaves, caffeine can enhance observably tea flavor [8].

Till date, the evaluation of tea quality level is still the taster's sensory assessment in the tea industry. It is performed by some trained tasters who have developed language of their own to describe various attributes of a tea infusion, and this language is sometimes difficult to comprehend by consumers. Additionally, the result by sensory evaluation is often less coherence and less impartiality because of physical or physiological factors [9].

Discrimination of quality level of tea can be achieved according to several analytical tools. Near-infrared reflectance spectroscopy [10], electronic nose [11] and electronic tongue [12] have been proposed. Nevertheless, the classification of tea varieties according to chemical composition has provided excellent results. The volatile components [13], free amino acids [14] and metal content [15–18] have been used as chemical parameters to differentiate tea varieties. Till now, the chemical compositions in green tea leaves have not been used for identification of tea quality level.

* Corresponding author. Tel.: +86 511 8790318 fax: +86 511 8780201.
E-mail address: q.s.chen@hotmail.com (Q. Chen).

0731-7085/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2008.09.016

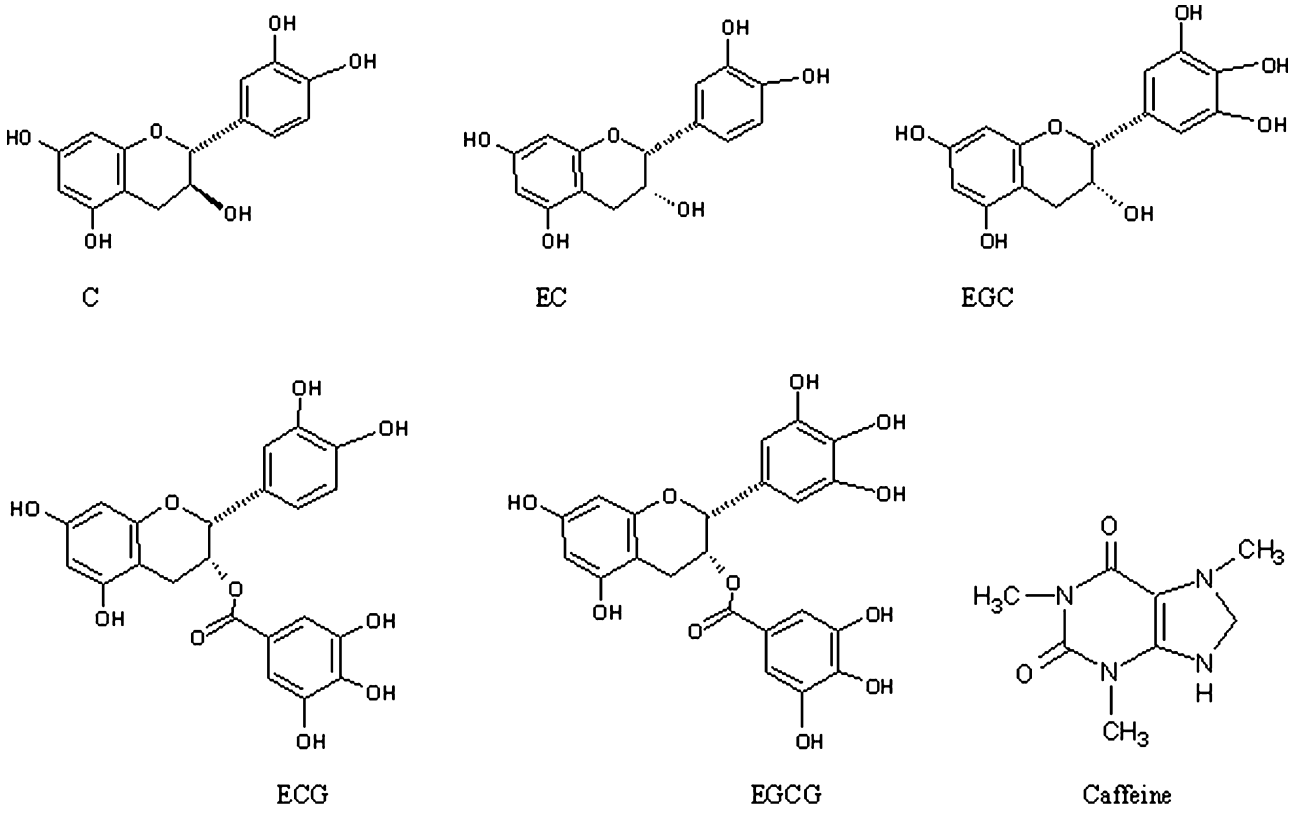

Fig. 1. Chemical structures of the catechins and caffeine.

In this work, five main catechins (EGCG, EGC, ECG, EC, and C) and caffeine in green tea leaves were measured simultaneously by HPLC. Their contents were used as chemical descriptors to identify the quality level of green tea. As a new chemical pattern recognition tool, support vector classification (SVC) was applied to identification of green tea's quality level in this work. SVC is an important branch in the statistical theory of support vector machine (SVM). SVC is incomparable to other pattern recognition algorithms. It is a state-of-the-art classification technique, which has a good theoretical foundation in statistical learning theory. SVC fixes the classification decision function based on the structural risk minimization (SRM) instead of the traditional empirical risk minimization (ERM), therefore, SVC pattern recognition avoid over-fitting problem [19,20].

2. Materials and methods

2.1. Standards and other chemicals

All standards viz. (+)-C, (−)-EC, (−)-EGC, (−)-ECG, and (−)-EGCG were purchased from the Sigma Chemical Co., USA. Caffeine was purchased from the Alfa-Aesar Chemical Co., USA. Acetonitrile, methanol and acetic acid (all HPLC grade) were purchased from Sinopharm Chemical Reagent Co. Ltd. (Shanghai, China). Other reagents were all analytical grade.

2.2. Sample preparation and preprocessing

Four grades of green tea were investigated in this work, which were purchased from the local supermarket in Zhenjiang City of China. All tea materials were already stored in the airtight containers within four months. Taking the heterogeneity of tea samples into consideration, major attention was paid to the sampling stage. The samples were ground before analysis. For the grinding, the whole tea leaves were put into a small electric coffee mill (Huanya, HY-02, Beijing, China) and ground for 10 s every time. After this procedure, the powders were sieved with a mesh width of 500 μm and the sieved powders were used for the subsequent analyses. Tea grade is often depicted tea quality as a comprehensive index. The former grade of tea often indicates higher quality of tea. In this work, the Grade 1 tea is the best quality, next, Grade 2, Grade 3 and Grade 4, in turn. As we know, the grade of tea is determined by some skillful tea tasters according to tea total quality including tea taste, aroma and tea appearances. In general, the internal attributes (including taste and aroma) of green tea are often the important references in determining its total grade. Internal attributes of green tea are often linked with many compositions in tea such as alkaloid and catechins, which are connected with planting geographical climate areas and the season of tea-leaf plucked.

In the experiment, 1.500 ± 0.001 g of tea powders were accurately weighted as a sample and each grade group had 15 samples, thus, 60 samples for 4 grades groups in all. Before analysis by HPLC, a sequence of sample preprocessing was implemented: each sample (1.500 ± 0.001 g of tea powders) is extracted twice with 80 mL of 70% aqueous methanol each for 30 min at a temperature of 80 °C. After cooling, the extracts are centrifuged at 3500 rpm for 10 min. The liquid phases of both extracts are collected in a 250 mL volumetric flask and made up to volume by 70% aqueous methanol. The tea brew is filtered through a 0.45 μm membrane filter, diluted 5 times with Millipore water and analyzed immediately.

2.3. Chromatographic conditions

To determine the content of catechins and caffeine, RP-HPLC method was applied in the Shimadzu LC-20A series (Shimadzu Co., Kyoto, Japan). The used column was a Shim-Pack VP-ODS Rx-C18

Table 1
Gradient elution system for the separation of the green tea catechins and caffeine (with mobile phase A: 9% acetonitrile, 2% acetic acid with 20 μg/mL EDTA; mobile phase B: 80% acetonitrile, 2% acetic acid with 20 μg/mL EDTA).

Time (min)	0	10	15	20	25
Mobile phase A (%)	85	85	75	75	85
Mobile phase B (%)	15	15	25	25	15

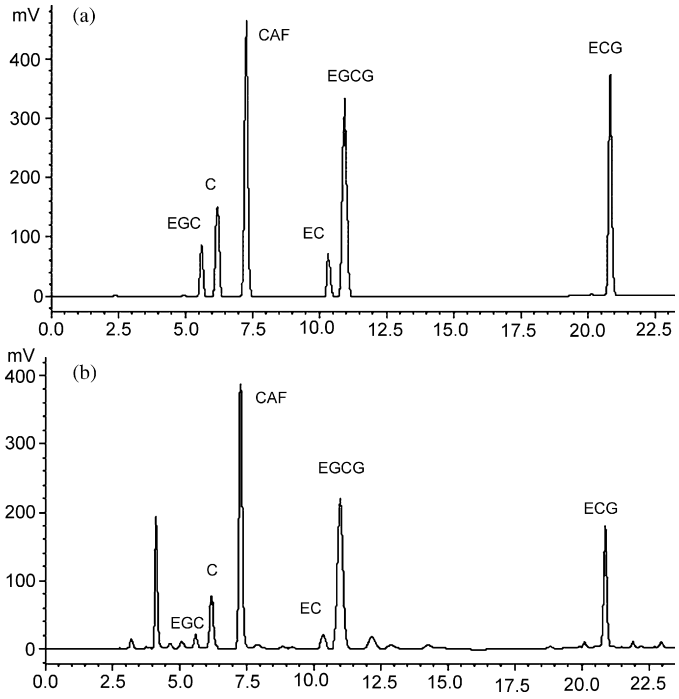

Fig. 2. HPLC separation of catechins and caffeine: (a) chromatogram of catechins and caffeine as calibrations standards; (b) chromatogram of tea sample.

column with 4.6 mm × 250 mm (i.d. × length) and 5 μm nominal particle size. The flow rate was set at 1.0 mL/min and the injected volume is 10 μL. The column temperature was kept at 35 ± 0.5 °C using a column oven. Gradient elution was applied in this work, which is shown in Table 1. Catechins separations are checked by the SPD-20A UV-detector, which was set at 278 nm. The HPLC separations of the catechins and caffeine are shown in Fig. 2.

2.4. Standards and calibration

Stock solutions of standards viz. five catechins and caffeine at 0.1 g/mL were prepared in 70% aqueous methanol and stored at −20 °C till further use. The standard solutions were passed through 0.45 μm membrane filter before injecting into HPLC. Standard curves for all standards were plotted by injecting 6 kinds of standard mixture and peak area responses were obtained. A standard graph was prepared by plotting concentration versus area. All peaks were plotted and integrated using software. The area and the retention time of the analyte peaks were compared with those of respective standards.

2.5. Support vector classification (SVC)

As a pattern recognition, support vector classification was attempted in this work. SVC is an important branch of support vector machine. SVM is based on statistical learning theory proposed by Vapnik and Chervonenkis [19,20]. Here, a brief introduction is presented, and readers can refer to the references and tutorials about SVM in detail [21,22]. The SVC is originated from the classification of two-class problems, in which SVC can be considered to create a hyperplane between two sets of data for classification. In case of 2-dimensional situation, The SVC tries to set an appropriate boundary so that the distance between the boundary and the nearest data point is maximal. The boundary is then placed in the middle of this margin. The nearest points that are used to define the margins are known as support vectors. Once the support vectors are selected, the rest of the feature set can be discarded, because the support vectors contain all the necessary information for the classifier.

In case the linear boundary in the low dimension input space would not be enough to separate two classes properly, it is possible to create a hyperplane that allows linear separation in the higher dimension feature space. In SVC, it is achieved through a transformation function $\Phi(x)$ that converts the data from a low dimension (n-dimension) input space to a high dimension (ε-dimension) feature space:

$$s = \Phi(x) \tag{1}$$

where $x \in R^n$ and $s \in R^\varepsilon$. Fig. 3 shows the transformation from the input space to the feature space, in which the non-linear boundary in the input space has been transformed into a linear boundary in the feature space. The transformation into higher dimensional space is implemented by introducing a kernel function. In brief, kernel function can map the data which are not linearly separable in input space into a higher dimensional feature space, where becomes linearly separable.

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \tag{2}$$

In SVM statistics theory, there are three classical kernel functions, which are polynomial kernel function, Gaussion kernel function, and sigmoid kernel function, respectively. Selection of kernel function has a high influence on the performance of SVC. In general, Gaussion kernel function is the optimal choice, without

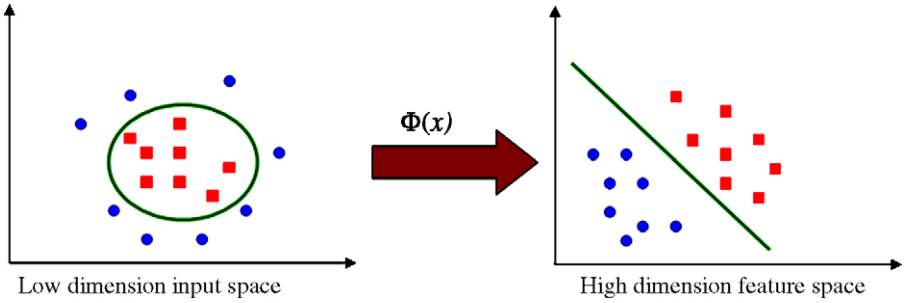

Fig. 3. Non-linear separation case in the low dimension input space and linear separation case in the high dimension feature space. Squares and circles denote the negative and the positive samples, respectively.

Table 2
Contents (% w/w, dry base) of catechins and caffeine in green tea samples.

Grades	S.N.	Parameters	EGC	C	CAF	EC	EGCG	ECG
Grade 1	15	Range	2.813–4.794	0.981–1.400	2.501–3.653	0.720–1.338	5.867–10.62	1.337–4.060
		Mean	3.359	1.144	3.178	0.979	9.303	3.468
		S.D.	0.461	0.129	0.375	0.147	1.328	0.755
Grade 2	15	Range	2.543–4.067	0.658–1.198	2.199–3.473	0.757–1.398	5.962–9.747	1.932–3.974
		Mean	3.261	0.916	2.892	0.977	8.448	3.332
		S.D.	0.398	0.164	0.388	0.165	0.944	0.512
Grade 3	15	Range	2.969–3.692	0.790–1.210	2.420–3.414	0.733–1.026	7.790–10.414	1.688–3.771
		Mean	3.423	1.004	2.894	0.947	9.365	2.569
		S.D.	0.201	0.117	0.301	0.078	0.808	0.850
Grade 4	15	Range	3.669–4.530	0.726–1.106	2.166–2.841	0.720–1.610	8.242–9.153	1.022–2.022
		Mean	4.037	0.853	2.514	0.897	8.619	1.639
		S.D.	0.292	0.126	0.193	0.211	0.260	0.245

prior experienced knowledge. The structure of Gaussion kernel function is the radial basic function (RBF), so it is also called RBF-kernel function. Gaussion kernel function is shown in Eq. (3).

$$K(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2K^2}\right) \quad (3)$$

where K is the kernel parameter, which is the bandwidth of the RBF function. The structure of Gaussion kernel is simple and fast computation compared with other kernel functions. Therefore, only Gaussion kernel was used in this work.

3. Results and discussion

In this work, 60 tea samples were attempted, and each grade has 15 samples. Six compounds contents of samples from each grade were measured simultaneously by HPLC as shown in Table 2. They were used to identify the quality level of green tea as six chemical descriptors.

The ultimate aim of this work is to identify four grades of tea by six chemical descriptors and supervised pattern recognition. Supervised pattern recognition refers to the techniques, in which classification model is developed on a training set of samples with categories. The model performance is evaluated by the use of a prediction set. Therefore, before building an identification model, all samples were divided into two subsets. One of subset was called the training set which was used to build model, and another was called the prediction set which was used to test the robustness of model. The training set contained 40 samples, and each grade had 10 samples. The remaining 20 samples constituted the prediction set, and each grade had 5 samples.

3.1. SVC model and identification results

To obtain a good performance, some parameters in SVC have to be optimized by cross-validation. These parameters include: (1) Regularization parameter (R), which determines the tradeoff between minimizing the training error and minimizing model complexity; (2) Kernel parameter (K), which is the bandwidth of the RBF-kernel function.

In this work, the kernel parameter (K) is arranged from 0.1 to 2 with 0.1 interval, and the regularization parameter (R) is arranged from 10 to 100 with 10 intervals. Thus, 20 K values (K = 0.1, 0.2, . . ., 2.0) and 10 R values (R = 10, 20, . . ., 100) are optimized simultaneously by cross-validation. The optimal model was searched by the highest recognition results by cross-validation as shown in Fig. 4. It can be found that the optimal SVC model is achieved when R = 60 and K = 1. The recognition rate by cross-validation is 95% for the optimal SVC model.

Table 3 shows the confusion matrix for the recognition results in the training and prediction sets. In the training set, the recognition results by cross-validation are obtained as follows: one Grade 2 sample is classified wrong to the Grade 3 group; one Grade 3 sample is classified wrong to the Grade 2 group; the remaining samples are all classified correctly; total recognition rate is 95%. In the prediction set, when the optimal SVC is used to test the new samples, the recognition results are obtained as follows: one Grade 2 samples are classified wrong to the Grade 3 group and one Grade 3 sample is classified wrong to the Grade 2 group; the remaining samples are all classified correctly; total recognition rate is 90%. The recognition results demonstrated that SVC model has high and robust recognition performance.

3.2. Discussion of identification results

Investigated from Table 3, misclassification often occurs between Grade 2 group and Grade 3 group, due to their close quality. Tea fresh leaves are often picked from late March and early April to July every year. Green tea processed with fresh leaves picked in early stage is regard as high quality level and its quality level declines with the passage of time. As we known, tea with the same category might share some similar sensory attributes, but there are subtle differences in chemical compounds such as catechins and caffeine. Picking time of tea fresh leaves has a few effects on the catechins and caffeine contents. Picking time from Grade 2 sample is close to Grade 3 sample, thus, their catechins and caffeine contents are very similar. These might explain the fact that misclassification often occurs between Grade 2 group and Grade 3 group.

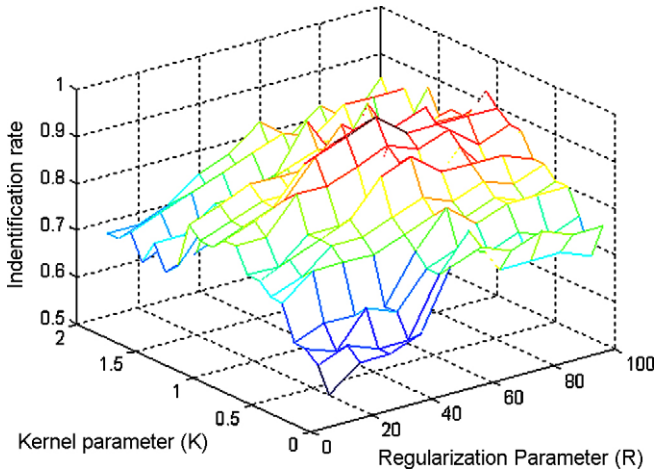

Fig. 4. Identification rates in different kernel parameter (K) and regularization parameter (R) by cross-validation.

Table 3
Confusion matrix for the recognition results in the training and prediction set.

Subset	Grades	Sample numbers	Recognition results				Total recognition rate
			Grade 1	Grade 2	Grade 3	Grade 4	
Training set	Grade 1	10	10	0	0	0	95%
	Grade 2	10	0	9	1	0	
	Grade 3	10	0	1	9	0	
	Grade 4	10	0	0	0	10	
Prediction set	Grade 1	5	5	0	0	0	90%
	Grade 2	5	0	4	1	0	
	Grade 3	5	0	1	4	0	
	Grade 4	5	0	0	0	5	

Table 4
Comparison of recognition results from SVC, BP-ANN and LDA models.

Models	Recognition results of models	
	Training set	Prediction set
SVC	95%	90%
BP-ANN	90%	75%
LDA	85%	80%

To highlight the good performance of the SVC algorithm, we attempted to compare SVC algorithm with linear discriminant analysis (LDA) and back propagation artificial neural network (BP-ANN) approaches in this work. Table 4 shows the recognition results obtained by LDA, BP-ANN and SVC approaches in the training and prediction set. As shown in Table 4, identification results by BP-ANN (95%) were equal to SVC (95%) approaches in the training set, and were better than recognition results of LDA (85%) approach. However, when predicted, the recognition result by SVC (90%) approach was obviously better than the results by BP-ANN (75%) and LDA (80%) approaches in the prediction set. Such phenomena might be explained by statistical learning theory.

SVC and BP-ANN are non-linear statistical learning approaches, while, LDA is a linear method. In general, non-linear method is stronger than linear method in the ability of self-learning and self-adjust. Therefore, identification results by BP-ANN and SVC approaches are better than identification results by LDA approach in the training set. Traditional neural network approaches including BP-ANN are based on the empirical risk minimization principle. It suffers difficulties with producing an over-fit model. The 'best' model by training often results in worse predictive result. In other words, the generalization of the model is worse. The foundation of SVC embodies the structural risk minimization principle, which has been shown to be superior to the ERM principle. SRM minimizes an upper bound on the expected risk, as opposed to ERM that minimizes the error on the training data [19–21]. Therefore, SVC shows its excellent generalization in identification results, and it also results in the better results than BP-ANN approach in the experiment.

4. Conclusions

HPLC was attempted to measure simultaneously five main catechins and caffeine in green tea. Their contents were used as chemical descriptors to identify the quality level of green tea. SVC as a pattern recognition tool was applied to develop identification model. The overall results sufficiently demonstrate that HPLC coupled with SVC pattern recognition can be identified tea quality level. In contrast to BP-ANN and LDA approaches, SVC algorithm shows its excellent performance in identification results. It can be concluded that HPLC analysis with SVC pattern recognition has a high potential to identification other agricultural product quality level.

Acknowledgements

This work has been financially supported by the National Natural and Science Foundation of China for Youth Program (Grant No. 30800666), the Natural Science Foundation for Colleges and Universities in Jiangsu Province (Grant No. 08KJB550003), and the Advanced Talents Science Foundation of Jiangsu University (Grant No. 08JDG007). We are also grateful to our colleagues for many stimulating discussion in this field.

References

[1] C.S. Yang, P. Maliakal, X. Meng, Annu. Rev. Pharmacol. Toxicol. 42 (2002) 25–54.
[2] K. Nakachi, S. Matsuyama, S. Miyake, M. Suganuma, K. Imai, BioFactors 13 (2000) 49–54.
[3] V.W. Setiawan, Z.F. Zhang, G.P. Yu, Q.Y. Lu, Y.L. Li, M.L. Lu, M.R. Wang, C.H. Guo, S.Z. Yu, R.C. Kurtz, C.C. Hsieh, Int. J. Cancer 92 (2001) 600–604.
[4] K. Shibata, M. Moriyama, T. Fukushima, A. Kaetsu, M. Miyazaki, H. Une, J. Epidemiol. 10 (2000) 310–316.
[5] L. Jian, L.P. Xie, A.H. Lee, C.W. Binns, Int. J. Cancer 108 (2004) 130–135.
[6] H. Fujiki, M. Suganuma, S. Okabe, E. Sueoka, N. Sueoka, N. Fujimoto, Y. Goto, S. Matsuyama, K. Imai, K. Nakachi, Mutat. Res. 480–481 (2001) 299–304.
[7] L.N. Seetohul, M. Islam, W.T. O'Hare, Z. Ali, J. Sci. Food Agric. 86 (2006) 2092–2098.
[8] Q.S. Chen, J.W. Zhao, H.D. Zhang, X.Y. Huang, X.Y. Wang, Anal. Chim. Acta 572 (2006) 77–84.
[9] Q. Chen, J. Zhao, J. Cai, J. Trans. ASABE 52 (2008) 623–628.
[10] Q.S. Chen, J.W. Zhao, C.H. Fang, D.M. Wang, Spectroc. Acta Pt. A 66 (2007) 568–574.
[11] H. Yu, J. Wang, H. Zhang, Y. Yu, C. Yao, Sens. Actuators B 128 (2008) 455–461.
[12] Q. Chen, J. Zhao, S. Vittayapadung, Food Res. Int. 41 (2008) 500–504.
[13] N. Togari, A. Kobayashi, T. Aishima, Food Res. Int. 28 (1995) 495–502.
[14] A. Alcazar, O. Ballesteros, J.M. Jurado, F. Pablos, M.J. Martin, J.L. Vilches, A. Navalon, J. Food Chem. 55 (2007) 5960–5965.
[15] P.L. Fernández-Cáceres, M.J. Martín, F. Pablos, A.G. González, J. Agric. Food Chem. 49 (2001) 4775–4779.
[16] M.Á. Herrador, A.G. González, Talanta 53 (2001) 1249–1257.
[17] P.L. Fernández, F. Pablos, M.J. Martín, A.G. González, Food Chem. 76 (2002) 483–489.
[18] A. Moreda-Piñeiro, A. Fisher, S.J. Hill, J. Food Compos. Anal. 16 (2003) 195–212.
[19] V.N. Vapnik, A.Y. Chervonenkis, Theory Probab. Appl. 17 (1971) 264–280.
[20] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[21] J.W. Zhao, Q.S. Chen, X.Y. Huang, C.H. Fang, J. Pharmaceut. Biomed. Anal. 4 (2006) 1198–1204.
[22] S.R. Gunn, Support Vector Machines for Classification and Regression, Technical Report: Image Speech and Intelligent Systems Research Group (Paper available on http://www.isis.ecs.soton.ac.uk/resources/svminfo/) (1998).