International Conference
AUTOMATICS AND INFORMATICS'15
4-7 October 2015, Sofia, Bulgaria

UV-VIS SPECTROSCOPY AND STATISTICAL CLUSTER ANALYSIS IN DISTINGUISHING DIFFERENT TYPES OF HONEY

D. Tsankova[1], S. Lekova[2]

[1]*University of Food Technologies – Plovdiv, 26 Maritsa Blvd, 4000 Plovdiv, Bulgaria, e-mail: dtsankova@yahoo.com*
[2]*University of Chemical Technology and Metallurgy – Sofia, 8 St. Kliment Ohridski Blvd, Sofia, e-mail: sv_lekova@uctm.edu*

Abstract: The aim of the article is to investigate the potential of honey discrimination (on the base of its botanical origins) by UV-Vis spectroscopy and subsequent statistical cluster analysis. Thirty-six samples from three types of honey (acacia, linden, and honeydew) are measured by a spectrophotometer "Cary100" with recorded wavelength range of 190~900 nm for calibration of honey classifier. Firstly, principal components analysis (PCA) is used for reducing the number of inputs (wavelengths) and a proper visualization of the experimental results. Next, the first two principal components (PCs) are combined separately with Naïve Bayes classification (NBC) and k-means clustering (KMC) to develop PC-NBC and PC-KMC models. The high accuracy of the proposed honey classifiers is confirmed by leave-one-out cross-validation test conducted in MATLAB environment.

Key words: UV-Vis spectroscopy, honey discrimination, PCA, Naive Bayes classification, K-means clustering

INTRODUCTION

"Honey is the natural sweet substance, produced by honeybees from the nectar of flowers or from secretions of living parts of plants or excretions of plant sucking insects on the living parts of plants, which the bees collect, transform by combining with specific substances of their own, deposit, dehydrate, store and leave in honeycombs to ripen and mature" [1, 2]. Honey consists of sugars, water, amino acids, oil, mineral salts and especial enzymes produced by bees [3].

Production of natural honey is a laborious process, which is time consuming and involves a lot of cost. Therefore honey is often subject to falsification by adding sugar and other impurities. Furthermore, some types of honey have a higher market price than others, so in order to prevent fraud in the labeling, it should be developed a means of distinguishing between different types of honey.

Several methods have been used for the determination of the floral origin of honey and among them the pollen recognition and sensory analysis are the most popular ones. However the technique of analysis of honey's pollen content is tedious and has some limitations. The other methods are mainly based on the analysis of honey's aroma compounds, sugar profile, flavonoid pattern, non-flavonoid phenolics, organic acids, isotopic relations, and protein and amino acid compositions and marker presence [4, 5]. But some of these methods are generally too time-consuming, complex, and labour intensive for routine quality control application or require very specialised personnel to interpret the results [6, 7].

In addition, most of the analytical techniques involve some kind of sample pre-treatment. The advantages of the technique of UV-visible (UV-Vis) and infrared (IR) spectroscopy with respect to other analytical methods are the non-invasive approach, the relatively easy and quick data acquisition. Some authors [8, 9] have used the IR technique for qualification of adulterants in honey with good accuracy. Recently, both near infrared (NIR) and middle infrared (MIR) spectroscopy, were successfully used for classification of unifloral and multifloral honeys [10, 11, 12]. Some authors have used Vis spectrometry for the same purpose (13, 14), but there is almost no information on the use of UV for classification of honey according to its botanical origin.

The purpose of the paper is to investigate the possibility of honey discrimination (based on its botanical origin) using UV-Vis spectroscopy in absorbance mode. Spectroscopic data obtained undergo subsequent statistical processing including: principal components analysis (PCA) for reducing the classifiers' number of inputs; Naïve Bayes classification (NBC) and k-means clustering (KMC) for cluster distinguishing. Technology of joint use of PCA and clustering techniques from classical type as NBC and KMC, not only contributes reduction of the area of the input data, but also overcomes some computational problems such as 'badly scaled or close to singular matrix'. The performance of the two calibration models is confirmed by leave-one-out-cross validation test in MATLAB environment.

MATERIALS AND METHODS

Honey spectrum acquisition. Thirty-five samples of three different types of honey (acacia – 11 samples; linden – 15 samples; and honeydew – 10 samples) were purchased from supermarkets (Lexie, Kaufland, Piccadilly) and from private producers. All samples of honey were diluted with distilled deionized water till 10% solution. The samples were annealed at room temperature (23-24^{0}C). The spectral characteristics of the honey were taken with a spectrophotometer Cary100 ranging from 190 to 900 nm at 1 nm sampling space.

Spectral readings of the three types of honey were treated with the aid of the following methods, as follows. First, the method of the PCs has reduced dimensionality of the input data, then they were classified using two classifiers – a supervisor Bayes classifier and unsupervisor k-means one.
Principal Components Analysis [15, 16]**.** The aim of the method is to reduce the dimensionality of multivariate data (e.g., wavelengths) whilst preserving as much of the relevant information as possible. PCA is a linear transformation, that

transforms the data (observations of possibly correlated variables) to a new coordinate system such that the new set of variables, *the principal components*, are linear functions of the original variables. PCs are uncorrelated, and the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on. This is achieved by computing the covariance matrix for the full data set. Then, the eigenvectors and eigenvalues of the covariance matrix are computed, and sorted according to decreasing eigenvalue [15, 16].

All the principal components are orthogonal to each other. The full set of principal components is as large as the original set of variables. Usually the sum of the variances of the first few principal components exceeds 80% of the total variance of the original data [17].

In this study, the first two PCs are used by the following two calibration methods: NBC and KMC, for developing discrimination models.

Naive Bayes classification algorithm [17, 18, 19, 20]. The Naive Bayes classifier is fast and easy to implement. It is designed for use when features are independent of one another within each class, but it appears to work well in practice even when that independence assumption is not valid. It classifies data in two steps – *training* and *prediction* steps. In the *training step*, using the training samples, the method estimates the parameters of a probability distribution, assuming that features are conditionally independent given the class. During the prediction step, for any unseen test sample, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test sample according the largest posterior probability.

In this study the NBC algorithm uses normal (Gaussian) distribution. It is appropriate for features that have normal distributions in each class. The Naive Bayes classifier estimates a separate normal distribution for each class by computing the mean and standard deviation of the training data in that class.

K-means clustering [21, 22]. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Given an initial set of k means, the algorithm proceeds by alternating between two steps - *assignment* and *update* steps [21]. During the *assignment step* each observation is assigned to the cluster whose mean yields the least within-cluster sum of squares. Since the sum of squares is the squared Euclidean distance, this is the "nearest" mean. Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means. In the *update step* the new means are calculated and they have to be the centroids of the observations in the new clusters.

The algorithm has converged when the assignments no longer change. Since both steps optimize the within-cluster sum of squares, and there exists a finite number of such partitions, the algorithm must converge to a local optimum. There is no guarantee that the global optimum is found by k-means algorithm.

RESULTS AND DISCUSSION

Absorbance Spectra: Absorbance spectra of the three types of honey with wavelengths ranging from 190 to 900 nm are shown in Figure 1. The spectra present peaks at the band of 190 ~ 350 nm and low absorbance in the range above 760 nm.

After receiving the "raw" data from the measurements, they are subject to pre-treatment. In order to remove some apparent interference received data are limited from above by a predetermined value (AbsorbMax). The resulting absorbtion curves are smoothed by the method of creeping averaging, using the formula:

$$a_{i+l/2} = \frac{1}{l+1}\sum_{k=0}^{l} a_{i+k}\,, \qquad (1)$$

where l is the width of a linear filter accepting even-numbered values. All the spectra are similar in spectral shape and absorbance. Therefore, it is necessary to apply appropriate methods of multivariate analysis to distinguish honey.

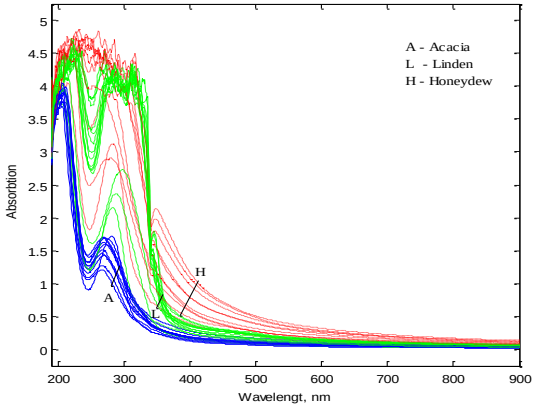

Figure 1. Absorbance spectra (acacia honey – solid line; linden honey – dashed line; honeydew honey – dotted line)

PC-NBC and PC-KMC Based Models for Honey Discrimination: The spectral dimensionality was reduced to a small number (two) of principal components using PCA. The scores scatter plot of the 1st and 2nd PCs is shown in Figure 2, where samples from classes 'acacia', 'linden' and 'honeydew' are marked with circular, triangular and squared symbols, respectively. It is evident that the samples form three clusters, that with a few exceptions coincide with the three types of honey mentioned above. The figure shows that one sample of linden honey is located in the cluster of acacia honey, and 4 samples (2 - linden honey and 2 - honeydew honey) are located far from their cluster centres in the space between three clusters. Here, determining the type of honey is based solely on the inscription on the label by the manufacturer, i.e. trusting the manufacturer. The first two PCs explain as high as 97.11 % of variance of the spectra (91.32 % for PC-1 and 5.79 % for PC-2). The two PCs were chosen to develop PC-NBC and PC-KMC models. One-out-cross-validation test was used to check the performance of the classifiers. The prediction results of the honey's botanical origin made by the proposed classifiers, PC-NBC and PC-KMC, are shown in Figure 3 / Table 1 and Figure 4 / Table 2, respectively. Table 3 shows the error prediction of the models mentioned above (PC-NBC and PC-KMC) for each class (acacia, linden and honeydew) separately and in total for the entire model. The performance of the PC-NBC based model is significantly better (97.22% accuracy) than the PC-KMC based one (86.11%) for honey discrimination.

CONCLUSION

Based on UV-Vis spectroscopy for discrimination of (botanical origins of) honey, two classification models, PC-NBC and PC-KMC, were developed for wavelengths ranging from 190 to 900 nm. As a result of using PCA the number of input data was reduced from 711 wavelengths to only 2 PCs. The obtained advantages are the lack of correlation between

the input data and also the ability to visualize the clusters formed by different types of honey. The PC-NBC and PC-KMC based classifiers show good prediction accuracies, 97.2% and 86.1%, respectively, determined by the 36 leave-one-out-cross-validation tests.

Future work will include: increasing the number of samples from the three classes mentioned above, adding new samples of other types of honey; using methods of artificial intelligence to increase the classifiers' accuracy; and etc.

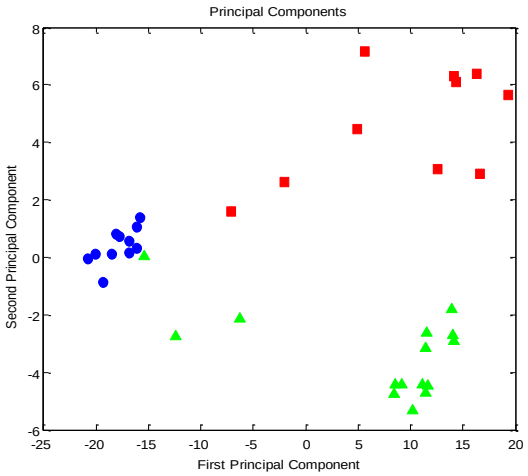

Figure 2. PCA of the three types of honey

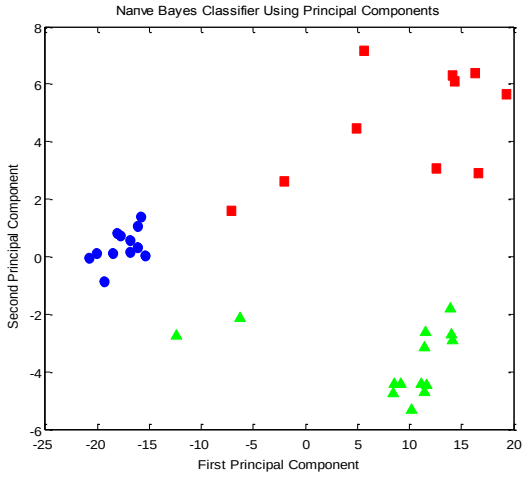

Figure 3. PC-NBC model of the three types of honey

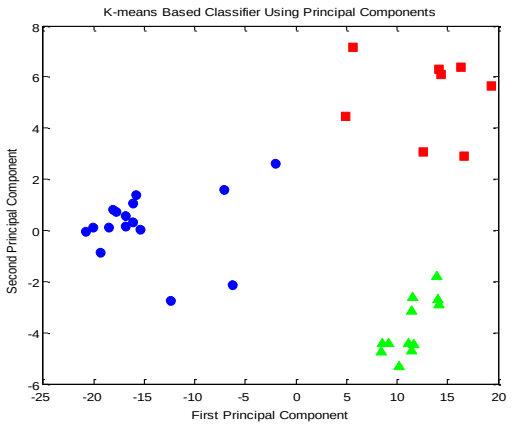

Figure 4. PC-KMC model of the three types of honey

Table 1: Discrimination accuracy of PC-NBC model

		Predicted Class by PC-NBC			
		Acacia	Linden	Honeydew	
Observed Class	Acacia	11	0	0	11
	Linden	1	14	0	15
	Honeydew	0	0	10	10
		12	14	10	

Table 2: Discrimination accuracy of PC-KMC model

		Predicted Class by PC-KMC			
		Acacia	Linden	Honeydew	
Observed Class	Acacia	11	0	0	11
	Linden	3	12	0	15
	Honeydew	2	0	8	10
		16	12	8	

Table 3: Discrimination accuracy of PC-NBC and PC-KMC models

Error prediction	PC-NBC (%)	PC-KMC (%)
Total	2.8	13.9
Acacia honey	9.1	45.4
Linden honey	6.7	20.0
Honeydew honey	0.0	20.0

ACKNOWLEDGEMENTS

The paper presents research and development, supported by Scientific Fund of Internal Competition of the University of Food Technologies – Plovdiv under the Research Project No.7/14-H.

REFERENCES

1. *Codex Alimentarius*: Draft revised standard for honey (at step 10 of the Codex procedure). Alinorm 01/25, pp.19-26, 2001.

2. *EU Council*: Council directive 2001/110/EC of 20 December 2001 relating to honey. Official Journal of the European Communities L10, pp.47–52, 2002.

3. Boffo, E.F., L.A. Tavares, A.C.T. Tobias, M.M.C. Ferreira, A.G. Ferreira. Identification of components of Brazilian honey by ^{1}H NMR and classification of its botanical origin by chemometric methods, LWT - Food Science and Technology, 49, pp.55-63, 2012, www.elsevier.com/locate/lwt

4. Anklam E., A review of the analytical methods to determine the geographical and botanical origin of honey. Food Chem., 63, pp.549-562, 1998.

5. Hermosin, I., R.M. Chicon, M.D. Cabezudo. Free amino acid composition and botanical origin of honey, Food Chemistry 83, pp. 263–268, 2003, www.elsevier.com/locate/foodchem

6. Bogdanov S., P. Martin. Honey authenticity, Mitt. Geb. Lebensmittelunters Hyg. 93, pp. 232–254, 2002.

7. Benedetti, S., S. Mannino, A.G. Sabatini, G.L. Marcazzan. Electronic nose and neural network use for the classification of honey, Apidologie 35, 2004, 1–6, © INRA/DIB-AGIB/ EDP Sciences, DOI: 10.1051/apido:2004025

8. Gallardo-Velazquez, T., G. Osorio-Revilla, M.Z.D. Loa, and Y. Rivera-Espinoza. Application of FTIR-HATR

spectroscopy and multivariate analysis to the quantification of adulterants in Mexican honeys, Food Research International, vol. 42, no.3, pp. 313–318, 2009.

9. Zhu, X., S.Li, Y. Shan et al. Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics, Journal of Food Engineering, vol.101, no. 1, pp. 92–97. 2010.

10. Davies A., B. Radovic, T. Fearn, E. Anklam. A preliminary study on the characterisation of honey by near infrared spectroscopy, J. Near Infrared Spectrosc. 10, pp.121–135, 2002.

11. Ruoff K., W. Luginbühl, S. Bogdanov, B. Estermann, T. Ziolko, R. Amado. Potential of Near Infrared Spectroscopy for Authenticity Testing of Unifloral Honey, Eur. Congr. for Authenticity of Food, Nyon, Switzerland, 2003.

12. Lichtenberg-Kraag B. Infrared spectroscopy: The quality assurance in honey analysis, Apidologie 34, pp. 479–480, 2003.

13. Li, Y., H. Yang. Honey Discrimination Using Visible and Near-Infrared Spectroscopy, International Scholarly Research Network, ISRN Spectroscopy, Volume 2012, Article ID 487040, 4 pages, 2012, doi:10.5402/2012/487040

14. Tsankova, D., S. Lekova. Botanical Origin-Based Honey Discrimination Using Vis-NIR Spectroscopy and Statistical Cluster Analysis, Journal of Chemical Technology and Metallurgy (JCTM), At Press.

15. Hotelling, H., Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 24(6 & 7), pp. 417–441 & pp. 498–520, 1933.

16. Jolliffe, I. T., Principal Component Analysis. Second ed. Springer Series in Statistics. New York: Springer-Verlag New York, 2002.

17. Statistics Toolbox™ User's Guide, R2014a, © COPYRIGHT 1993–2014 by The MathWorks, Inc.

18. http://en.wikipedia.org/wiki/Naive_Bayes_classifier#cite_note-rennie-3

19. McCallum, A., K. Nigam. "A comparison of event models for Naive Bayes text classification". AAAI-98 workshop on learning for text categorization 752, 1998.

20. Rennie, J., L. Shih, J. Teevan, D. Karger. "Tackling the poor assumptions of Naive Bayes classifiers". ICML, 2003.

21. MacKay, D., "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284–292. ISBN 0-521-64298-1. MR 2012999, 2003.

22. http://en.wikipedia.org/wiki/K-means_clustering#Independent_component_analysis_.28ICA.29