

Automatic Classification of Herbal Substances Enhanced with an Entropy Criterion

Victor Mendiola-Lau¹(✉), Francisco José Silva Mata¹, Yoanna Martínez-Díaz¹,
Isneri Talavera Bustamante¹, and Maria de Marsico²

¹ Advanced Technologies Application Center, Havana, Cuba
{vmendiola,fjsilva,ymartinez,italavera}@cenatav.co.cu

² Università degli Studi de Roma, Rome, Italy
demarsico@di.uniroma.it

Abstract. This paper presents a novel automatic pattern recognition system for the classification of herbal substances, which comprises the analysis of chemical data obtained from three analytical techniques such as Thin Layer Chromatography (TLC), Gas Chromatography (GC) and Ultraviolet Spectrometry (UV), composed of the following stages. First, a preprocessing stage takes place that ranges from the TLC plate image conversion into a spectrum to the normalization and alignment of spectral data for all techniques. Then, a hierarchical clustering procedure is applied for each technique with the goal of discovering groups or classes that provide evidence concerning the different existing types. Next, an entropy-based template selection step for each group was introduced to exclude the less significant samples, thus allowing to improve the quality of the training set for each technique. In this manner, each class is now described by a set of key prototypes that allows the field expert to have a more accurate characterization and understanding of the phenomenon. Moreover, an improvement of the computational complexity for training and prediction tasks of the Support Vector Machines (SVM) is also achieved. Finally, a SVM classifier is trained for each technique. The experiments conducted show the validity of the proposal, showing an improvement of the classification results on each technique.

Keywords: Herbal substance · Clustering · Entropy · Template selection

1 Introduction

The identification of types of substances is a common task in analytical chemistry. This process entails not only recognizing types of substances or its components, but also discovering features that allow to distinguish among them. These differences arise commonly from types of processing, growing conditions, and geographic location. Recently, in analytical chemistry, the application of computational methods in the analysis of data resulting from analytical techniques, has allowed to increase the speed and accuracy of the system response, thus assisting the field expert in the decision taking process [7]. Analytical techniques such as

Thin Layer Chromatography (TLC) provide a very economical and straightforward choice to identify chemicals, such as plant extracts or mixtures. However, in order to reach safer conclusions, this technique is usually used in combination with other analytical techniques such as Ultraviolet Spectrometry (UV) and Gas Chromatography (GC) with the purpose of combining information and thus, having a more robust criterion to issue a verdict on the identity of an unknown substance or its type. The identification and classification of herbal substances is a complex problem given the fact that these are composed of a mix of many active compounds, which is where our study is focused. In this paper, we use Cannabis samples as representatives of these types of substances, though our proposal can be applied to any kind of mixed substances.

To our knowledge, no work in the literature proposes an automatic system able to integrate all the process steps: from data acquisition arising from several chemical analytical techniques, to obtaining suitable classification models that allow a proper characterization of the chemical class composition, based on the information provided by each technique. Figure 1 summarizes the main steps of our proposal for a completely automatic identification process, which represents a step forward in the solution of these complex problems regarding herbal substances identification.



Fig. 1. Proposed methodology for the identification process.

The paper is organized as follows. Section 2 describes the processing applied to spectral data. Next, Sect. 3 discusses the class discovery strategy. Section 4 proposes entropy-based criterion used to select the best templates. In Sect. 5, the classification procedure is presented. Discussion of experimental results is carried at Sect. 6. Conclusions and future work are outlined in Sect. 7.

2 Data Preprocessing

Acquired data for UV and GC analytical techniques have a spectral nature (see Fig. 2 (center) and (right)). However, in the case of the TLC technique, each sample is obtained in an image as spots linearly distributed on a narrow band or lane. Each spot corresponds to the presence of an active compound of the mix. After a photometric normalization of the image lane, a projective integral operation was applied to obtain a spectrum as shown in Fig. 2 (left). This new spectral representation for the TLC technique is more compact and robust than an image and at the same time, the field expert can analyze altogether the three techniques at once.

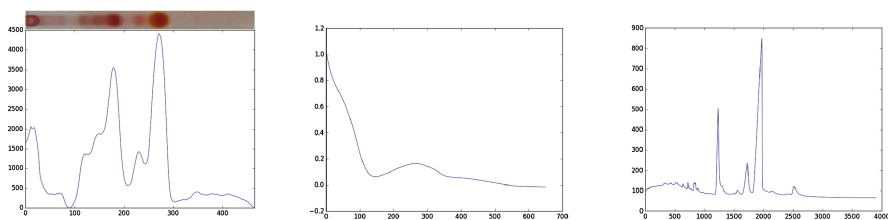


Fig. 2. The corresponding spectra of a substance for different analytical techniques: TLC (left), UV (center) and GC (right).

Next, high frequency and low amplitude noise are removed to prevent them from being confused with real peaks of substances components. Smoothing the spectra through spline interpolation [8] (see Fig. 3) was used. Chromatographic methods (TLC and GC) introduce experimental or instrumental errors. These so-called artifacts are mainly shifted peaks, co-elution of peaks, background offset, displacement of the baseline and scaling effects [9]. For these reasons, an alignment process is needed, which consists of adjusting the correspondence of concentration peaks of each substance, in accordance with their retention factors, represented as time. This process must be performed with respect to a suitable reference spectrum selected for each technique. The main selection criterion was to use the pattern that best correlates with others patterns, i.e. the one providing the maximum cumulative product of correlation coefficients. As a secondary criterion, the presence of peaks corresponding to basic chemical components of the substances is taken into account [9]. In order to deal with the problems of scaling and baseline shifts, the Multiplicative Scatter Correction (MSC) was employed [4]. Classical alignment algorithms such as the Correlation Optimized Warping (COW) and the Dynamic Time Warping (DTW) [10] were evaluated, where the better results were achieved by COW.

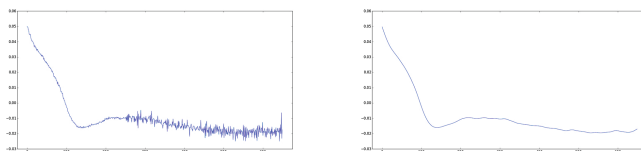


Fig. 3. Smoothing of the UV spectra by spline interpolation.

3 Discovery of Classes

Since the number of existing classes is unknown, an unsupervised clustering technique such as *hierarchical clustering analysis* (HCA) with a complete linkage criterion was applied. A metric to select the optimal partition was not employed, as the final partition was determined by means of an exploratory analysis assessed

by a field expert, where the presence and relation among the concentrations of active chemical components were analyzed. Given its precision and resolution capability, GC data was used as *ground truth* for the exploratory analysis performed on the UV and TLC data sets.

Two substances can be considered similar regarding the correlation of their composition, i.e., they have the same compounds present but also with a similar relative concentrations among them. For this reason, two correlation distances [6, 11] were selected as measures of similarity aiming to capture this behavior.

Figure 4 shows the clustering results of four of our classes, each class having more or less the same amount of samples for the same technique. The clusters labeled as 1 and 4 agreed in a 100 % of samples for the three techniques. Similarly, clusters labeled as 2 and 3 agreed only in 90 % and 75 % of the samples respectively. Moreover, it can be noticed that each cluster, regardless of the analytical technique it belongs to, contains spectra with similar shape.

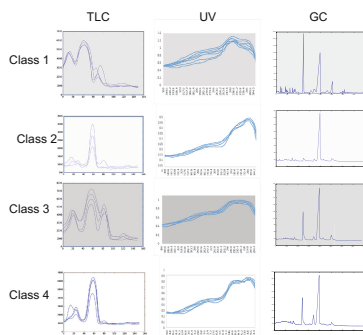


Fig. 4. Visual inspection of our final clustering results.

4 Entropy-Based Template Selection

In many works, models derived and somehow related with Information Theory have been used. Fields range from information communication to biometrics. Particularly, strategies for template selection based on the concept of entropy have been proposed [2], achieving very good results in identifying the most representative templates in a given set. The core concept relies on the computation of the entropy of a gallery of templates, which can be useful for the template selection task. This concept can be extended to our scenario, where each labeled group is considered as class of a given substance.

Given a gallery G of templates, this can be considered as the union of sub-galleries G_k where every $g_{i,k} \in G_k$, with $i = 1, 2, \dots, |G_k|$, belong to the same class k . Therefore, $G = \cup_{k=1}^K G_k$ and $G_k \cap G_h = \emptyset, \forall k \neq h$. A similarity measure s associates a real scalar value to a pair of templates. The comparison of a probe v with a gallery template $g_{i,k}$ can be denoted by $s_{i,v} = s(v, g_{i,k})$ which

can be normalized to be a value in the real interval $[0, 1]$ [2]. If we had a distance measure or a dissimilarity coefficient $d_{i,v} = d(v, g_{u_i,k})$, it is possible to consider the similarity value as $s_{i,v} = 1 - d_{i,v}$. Assuming that template v was correctly assigned to the class k , each similarity value $s_{i,v}$ can be interpreted as the probability that template v conforms to $g_{i,k}$, that is $s_{i,v} = p(v \approx g_{i,k})$.

According to this interpretation, a proper normalization of $s_{i,v}$ is needed, to have it ranging in the interval $[0, 1]$. Also, $\sum_{i=1}^{|G_k|} s_{i,v} = 1$ holds given the fact that v was correctly assigned to G_k . It is now possible to compute the (probe-related) entropy of the probability distribution obtained applying (4) to the whole G_k with respect to a probe v as follows:

$$H(G_k, v) = -\frac{1}{\log_2 |G_k|} \sum_{i=1}^{|G_k|} s_{i,v} \log_2 (s_{i,v}), \quad (1)$$

where $\frac{1}{\log_2 (|G_k|)}$ is a normalization factor, corresponding to the maximum entropy obtained when (4) has the same value for all the templates in the gallery G_k . In this way, the value obtained for $G(G_k, v)$ is normalized in the range $[0, 1]$. Finally, the entropy for the gallery G_k is computed by considering each template $g_{j,t}$ in turn as a probe v :

$$H(G_k) = -\frac{1}{\log_2 |Q|} \sum_{q_{i,j} \in Q} s_{i,j} \log_2 (s_{i,j}), \quad (2)$$

where Q represents the set of pairs $q_{i,j} = (q_{i,k}, q_{j,k})$ of elements in G_k such that $s_{i,j} > 0$. Once again, values are in the range $[0, 1]$ irrespective of the size of the gallery. $H(G_k)$ represents a measure of heterogeneity for G_k and it is worth noticing that having a common range for $H(G_k)$ allows for comparison of the representativeness of different galleries. Given a gallery G_k , the computation of $H(G_k)$ can be used to select a subset of representative samples out of it. The procedure described in [2] is based on an ordering of the gallery templates according to a representativeness criterion.

After performing sample ordering according to the above procedure, several criteria may be used to select those templates that guarantee a suitable representativeness of the gallery. A possible criterion is the *one-shot* selection proposed in [3]. With this strategy, the entropy difference is computed between each template and the preceding one and all templates are sorted according to this difference value. Finally, the k first templates are chosen as the k best representative templates. Another suitable strategy in this direction could be to select the *top-percentage* of templates according to the entropy difference, which has the additional advantage of allowing to control the resulting data set size.

It is worth underlining that the same technique, depending on the application requirements at hand, can be used for either single sub-galleries G_k , as in the present case, or for an overall gallery G of templates.

5 SVM Classification

Partitions obtained after the selection of the most representative templates are used for training SVM classifiers, one for each technique. Two substances can be considered *similar*, when their components have similar relative concentrations. The standard correlation coefficient and Spearman correlation coefficient are commonly used to account for spectral similarity [1]. Due to the effectiveness of these correlation distances in determining the classes for training and as an effective measure of dissimilarity between the spectra, it was decided to use correlation kernels for the implementation of SVM classifiers [5]. For the correlation kernel based on the standard correlation coefficient, Eq. 3 and Eq. 4 were exploited, where \bar{X} is the sample mean of data matrix \mathbf{X} . On the other hand, for the correlation kernel based on the Spearman correlation coefficient, Eq. 3 and Eq. 5 [11] were exploited, where n corresponds to the amount of spectral bands and x_i, y_i correspond to the i -th spectral bands.

$$\text{corr}(\mathbf{X}) = \begin{pmatrix} \text{corr}(X_1, X_1) & \text{corr}(X_1, X_2) & \dots & \text{corr}(X_1, X_n) \\ \text{corr}(X_2, X_1) & \text{corr}(X_2, X_2) & \dots & \text{corr}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{corr}(X_n, X_1) & \text{corr}(X_n, X_2) & \dots & \text{corr}(X_n, X_n) \end{pmatrix}, \quad (3)$$

$$\text{corr}(X, Y) = \frac{(X - \bar{X})^T (Y - \bar{X})}{\sqrt{(X - \bar{X})(X - \bar{X})^T} \sqrt{(Y - \bar{X})^T (Y - \bar{X})}}, \quad (4)$$

$$\text{spear}(X, Y) = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{\sqrt{[n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2][n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2]}}, \quad (5)$$

The correlation kernels are based on the correlation matrix defined above and the kernel values for any X_i and X_j are defined as:

$$K_{\text{corr}}(X_i, X_j) = e^{(1 - \text{corr}(X_i, X_j))}, \quad K_{\text{spear}}(X_i, X_j) = e^{(1 - \text{spear}(X_i, X_j))}. \quad (6)$$

6 Experimental Results and Discussion

The data sets obtained from the proposed template selection strategies as well as the original ones, were used as training and validation sets in order to assess their descriptive power. Table 1 shows the best template selection strategies compared with the original data sets (used as baseline) for each analytical technique: GC (142 samples), UV (143 samples) and TLC (92 samples). Column 2 shows the template selection strategies analyzed for each technique. In Column 3, the size of optimized data sets yielded by each selection strategy is provided. Finally, Columns 4 and 5 show the classification accuracy (CA) for a 10-fold cross validation procedure obtained for the correlation measures used at the different stages of the process: class discovery, entropy-based template selection and building a precomputed kernel for the SVM classifier.

Our system obtained a very good classification accuracy for the template selection strategies analyzed, almost always improving the classification results

Table 1. Improvement on classification and training set reduction of our strategy over the original data sets.

Technique	Selection strategy	No. samples	CA (Correlation)	CA (Spearman)
GC	baseline	142	95.07 %	94.37 %
	one-shot	58	96.61 %	98.28 %
	top-25 %	38	92.11 %	92.11 %
	top-50 %	74	98.65 %	93.24 %
	top-75 %	111	97.30 %	95.50 %
UV	baseline	143	96.50 %	95.10 %
	one-shot	52	98.08 %	96.23 %
	top-25 %	38	92.11 %	89.47 %
	top-50 %	73	97.26 %	95.89 %
	top-75 %	110	97.27 %	96.36 %
TLC	baseline	92	96.74 %	89.13 %
	one-shot	80	98.75 %	96.77 %
	top-25 %	22	95.45 %	81.82 %
	top-50 %	44	97.73 %	93.18 %
	top-75 %	66	96.97 %	90.91 %

compared to those obtained for the original data sets, therefore showing the validity of the proposal. The only selection strategy that could not achieve better results was the selection of the 25 % most representative templates of each class (top-25 %), this is due to the fact too few templates were selected, thus excluding others with relevant information for the class description. As can be seen, an additional advantage of the proposed methodology is a significant reduction of the training sets, thus improving the computational efficiency in the training and classification processes. As shown, the *one-shot* variant is usually a very good starting point, but it is also advisable to apply a complementary strategy for template selection. In some cases, the best results were obtained for the *top-percentage* strategy, which not only selected the same templates than the one-shot strategy, but also other templates contribute to improve the classification accuracy. Also, it is important to stress that the measure for spectral comparison based on the standard correlation coefficient showed a higher performance.

Due to differences in cluster size, the *one-shot* criterion was modified to select templates with an associated entropy difference greater than a specific threshold, instead of choosing a fixed k per cluster. Under these conditions, this strategy depends on a threshold, which could result in excluding templates that contribute with valuable information or the opposite, selecting templates that alter the class description or contribute with redundant information. On the other hand, the *top-percentage* criterion does not suffer from this problem with cluster sizes, but it is still necessary to define an optimal percent of representative templates.

7 Conclusions and Future Work

The main contribution of this work is the introduction of a novel pattern recognition system designed to cope with the problem of herbal substances classification and identification by combining different analytical techniques. As previously stated, the entropy-based template selection strategy is beneficial not only from a classification enhancement point of view, but also for the great impact in the process of class description for the field expert. Also, the measure used for spectral comparison based on the standard correlation coefficient proved to be more suitable than the one based on the Spearman correlation coefficient. The results obtained show that the proposal achieved an outstanding classification accuracy, specially considering the complexity of the classification of herbal substances. Moreover, the size of the training sets and the computational effort necessary in the training and classification tasks was reduced significantly. As future work, a thorough analysis should be carried concerning the selection of an optimal threshold for the one-shot selection strategy and the optimal percentage of templates to choose per cluster. Also, additional studies regarding a suitable combination of the independent classifiers would be of great interest.

References

1. Bodis, L.: Quantification of spectral similarity: towards automatic spectra verification. Ph.D. thesis, Babes-Bolyai University (2007)
2. De Marsico, M., Nappi, M., Riccio, D., Tortora, G.: Entropy-based template analysis in face biometric identification systems. *Sig. Image Video Process.* **7**(3), 493–505 (2013)
3. De Marsico, M., Riccio, D., Vazquez, H.M., Calana, Y.P.: GETSEL: gallery entropy for template selection on large datasets. In: International Joint Conference on Biometrics (IJCB), pp. 1–8 (2014)
4. Helland, I.S., Næs, T., Isaksson, T.: Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometr. Intell. Lab. Syst.* **29**(2), 233–241 (1995)
5. Jiang, H., Ching, W.K.: Correlation kernels for support vector machines classification with applications in cancer data. *Comput. Math. Methods Med.* (2012)
6. Kumar, V., Chhabra, J.K., Kumar, D.: Performance evaluation of distance metrics in the clustering algorithms. *J. Comput. Sci.* **13**(1), 38–52 (2014)
7. Muda, A.K., Choo, Y.H., Abraham, A., Srihari, S.N.: Computational Intelligence in Digital Forensics: Forensic Investigation and Applications. Springer, Heidelberg (2014)
8. Schumaker, L.: Spline Functions: Basic Theory. Cambridge University Press, Cambridge (2007)
9. Skov, T.H.: Mathematical resolution of complex chromatographic measurements. Ph.D. thesis (2008)
10. Tomasi, G., Savorani, F., Engelsen, S.B.: icoshift: an effective tool for the alignment of chromatographic data (2011)
11. Zwillinger, D., Kokoska, S.: CRC Standard Probability and Statistics Tables and Formulae. CRC Press, Boca Raton (1999)