New Dissimilarity Measures for Ultraviolet Spectra Identification

Andrés Eduardo Gutiérrez-Rodríguez[1], Miguel Angel Medina-Pérez[1], José Fco. Martínez-Trinidad[2], Jesús Ariel Carrasco-Ochoa[2], and Milton García-Borroto[1,2]

[1] Centro de Bioplantas. Carretera a Morón km 9, Ciego de Ávila, Cuba
[2] Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro No. 1, Sta. María Tonanzintla, Puebla, México, C.P. 2840
{andres,migue,mil}@bioplantas.cu,
{fmartine,ariel}@ccc.inaoep.mx

Abstract. Ultraviolet Spectra (UVS) analysis is a frequent tool in tasks like diseases diagnosis, drugs detection and hyperspectral remote sensing. A key point in these applications is the UVS comparison function. Although there are several UVS comparisons functions, creating good dissimilarity functions is still a challenge because there are different substances with very similar spectra and the same substance may produce different spectra. In this paper, we introduce a new spectral dissimilarity measure for substances identification, based on the way experts visually match the spectra shapes. We also combine the new measure with the Spectral Correlation Measure. A set of experiments conducted with a database of real substances reveals superior results of the combined dissimilarity, with respect to state-of-the-art measures. We use Receiver Operating Characteristic curve analysis to show that our proposal get the best tradeoff between false positive rates and true positive rates.

Keywords: Ultraviolet Spectra, Ultraviolet Spectra Comparisons Functions, Substance Identification, Dissimilarity Measures.

1 Introduction

Ultraviolet Spectra (UVS) represent, for a given substance, the relation between ultraviolet light absorbance and light wavelength. Due to UVS are unique for each pure substance, they are frequently used for substance identification in different areas such as medicine, geology, criminalistics, and industrial applications [1–4].

Identifying substances by UVS is a challenge because there are different substances with very similar spectra shape (Fig. 1). Additionally, different concentrations of the same substance produce different spectra, dilated or contracted, according to Lambert-Beer Law [5] (Fig. 2).

In this paper, we focus on substance identification (mainly drugs, medicines, poisons, pesticides and other organic substances) by ranking its spectrum according to its dissimilarity values against the spectra in a database. These substances

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2010, LNCS 6256, pp. 220–229, 2010.

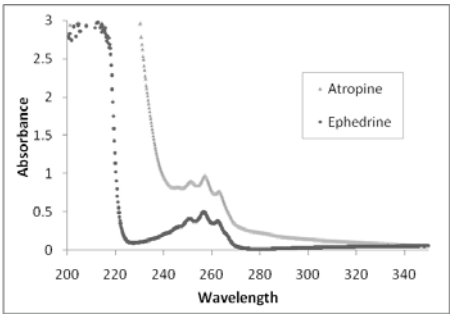

Fig. 1. Ultraviolet spectra of two different substances

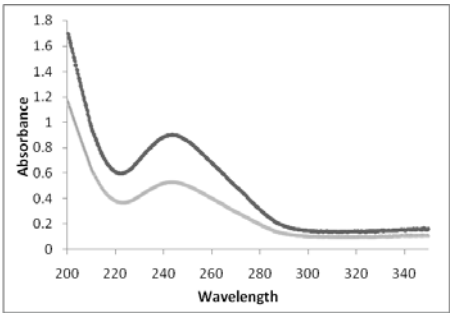

Fig. 2. Two spectra of Prednisone at different concentrations

are generally free of impurities or they have a predominant active chemical, which is involved in the identification process. In general, the quality of the dissimilarity function is the most important factor in the identification results.

In order to design a good UVS comparison function, we must take into account the application domain [1]; some of the most used properties are: agreements in the amplitudes of the signal, the shape of the spectrum, and a unique configuration of peaks that may change slightly. For substance identification, it is necessary a qualitative analysis of its spectrum. This leads to an empirical comparison of the unknown spectrum details with other known spectra. These details are maxima, minima and inflection points of the spectrum. Usually, experts visually match spectra based on their shape [6].

There are several UVS comparison functions [4, 7–12], but none of them effectively compare spectra shapes. The measures that compare spectrum absorbance values fail in comparing spectra of the same substance at different concentration, because these values can differ considerably. A normalization of the absorbance values introduces false positive matches when the spectrum has close absorbance values but differ in monotony or concavity. One attempt to overcome this problem is addressed in [1]. This measure compares spectra using the first derivatives, but ignores the changes in the curve concavity.

In this paper, we propose a new dissimilarity measure to compare ultraviolet spectra: Derivatives Sign Differences (DSD). DSD compares the spectral monotony and concavity, using the first and second derivatives of the 2D-shape of the spectrum at each wavelength. This way, the new measure can effectively match spectra of the same substance at different concentrations.

We compare DSD with several dissimilarity measures for substances identification, using a database containing 206 spectra of 103 substances (two spectra from each substance). This database was created by forensic experts in State Forensic Laboratory of Ciego de Avila. In order to evaluate the performance of the comparison functions we make use of Receiver Operating Characteristics (ROC) curves [13]. The experimental results show a good performance of the new measure, compared with eight UVS dissimilarities proposed in the literature. Moreover, we combine DSD with the Spectral Correlation Measure (SCM) [8], showing that this combination outperforms both single measures.

2 UVS Dissimilarities

There are several dissimilarity functions to compare mass spectra, infrared spectra, ultraviolet spectra, multi and hyper spectral images [11]. In this section, we briefly review some of the most cited spectral dissimilarities.

Perhaps, the most popular UVS measure is the Spectral Angle Mapper (SAM) [7]. SAM is primarily introduced for comparing hyperspectral image data; it compares two spectra by finding the angle between their absorbance tuples. Equation 1 shows SAM transformed to dissimilarity.

$$\mathrm{SAM}(s,t) = 1 - \left(\frac{\sum_{l=1}^{L} s_l t_l}{\sqrt{\sum_{l=1}^{L} s_l^2 \sum_{l=1}^{L} t_l^2}} + 1 \right) /2 \tag{1}$$

The tuple $s = (s_1, s_2, \ldots, s_l)$ represents a spectrum, where each s_l is the ultraviolet light absorbance for the corresponding light wavelength value w_l.

The main drawback of SAM is that the angle between tuples of two spectra of the same substance at different concentrations may be very different from zero. Van der Meer [8] proposes the Spectral Correlation Measure (SCM), which overcomes this limitation by standardizing the data, i.e. centralizing them using the mean of s and t [14].

$$\mathrm{SCM}\,(s,t) = 1 - \left(\frac{L \sum_{l=1}^{L} s_l t_l - \sum_{l=1}^{L} s_l \sum_{l=1}^{L} t_l}{\sqrt{\left[L \sum_{l=1}^{L} s_l^2 - \left(\sum_{l=1}^{L} s_l \right)^2 \right] \left[L \sum_{l=1}^{L} t_l^2 - \left(\sum_{l=1}^{L} t_l \right)^2 \right]}} + 1 \right) /2 \tag{2}$$

The Spectral Information Divergence (SID) [9] measures the information divergence between the probability distributions generated by two spectra. To do

so, SID models spectra as random variables by defining $p_k = s_k / \sum_{l=1}^{L} s_l,\ k = 1, \ldots, L$ so that $p = (p_1, \ldots, p_L)$. Then, it compares spectra taking into account the relative entropy between them:

$$\mathrm{SID}\,(s,t) = \sum_{l=1}^{L} p_l \log \frac{p_l}{q_l} + \sum_{l=1}^{L} q_l \log \frac{q_l}{p_l} \tag{3}$$

In a similar way as SAM, SID is introduced for comparing hyperspectral images.

SID-SAM is a combination of SID with SAM to enhance the spectral discriminatory probability [10]. The authors formulate this combination in two versions (Equation 4 and Equation 5).

$$\mathrm{SID} - \mathrm{SAM1}\,(s,t) = \mathrm{SID}\,(s,t) \sin\left(\arccos\left(1 - 2\mathrm{SAM}\,(s,t)\right)\right) \tag{4}$$

$$\mathrm{SID} - \mathrm{SAM2}\,(s,t) = \mathrm{SID}\,(s,t) \tan\left(\arccos\left(1 - 2\mathrm{SAM}\,(s,t)\right)\right) \tag{5}$$

Another fusion of spectral dissimilarities is the Spectral Similarity Scale (SSS) [11], which combines a modification of SCM with the Euclidean distance (Equation 6).

$$SSS\,(s,t) = \sqrt{1/L \sum_{l=1}^{L} \left(s_l - t_l\right)^2 + \left(1 - r\right)^2} \tag{6}$$

where:

$$r = \frac{\sum_{l=1}^{L} \left(s_l - 1/L \sum_{l=1}^{L} s_l\right) \cdot \left(t_l - 1/L \sum_{l=1}^{L} t_l\right)}{\sqrt{\sum_{l=1}^{L} \left(s_l - 1/L \sum_{l=1}^{L} s_l\right)^2 \cdot \sum_{l=1}^{L} \left(t_l - 1/L \sum_{l=1}^{L} t_l\right)^2}} \tag{7}$$

A significant drawback in the use of the Euclidean distance is that it is unbounded, because the range of values increases as the number of light wavelengths increases. Moreover, comparing spectra of the same substance at different concentrations (Fig. 2) would return high dissimilarity values. Robila and Gershman [12] propose the Normalized Euclidean Distance (NED) to overcome these limitations (Equation 8). The main difference with the Euclidean distance is that NED normalizes each spectrum dividing the absorbance values by the average absorbance.

$$\mathrm{NED}\,(s,t) = \sqrt{\sum_{l=1}^{L} \left(\frac{s_l}{1/L \sum_{l=1}^{L} s_l} - \frac{t_l}{1/L \sum_{l=1}^{L} t_l}\right)^2} \tag{8}$$

Fig. 3 shows the effects of normalizing spectra from Fig. 2. Notice that both spectra have now similar absorbance values. Nevertheless, there are examples where this type of normalization does not perform correctly. For example, Fig. 4 shows two spectra of the same substance after normalization with a clear difference between absorbance values for almost all wavelength values. In this case, NED erroneously returns a high dissimilarity value.

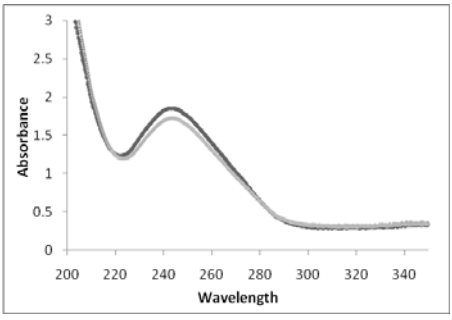

Fig. 3. Spectra from Fig. 2 normalized using the average of the absorbance values

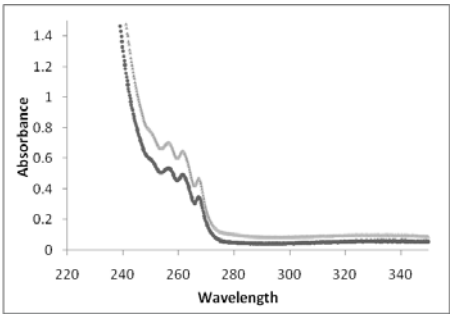

Fig. 4. Two spectra of Ampicillin after normalization

Paclík and Duin [1] propose to incorporate the difference between shapes of spectra into the dissimilarity measure. For that purpose, they use the spectra first derivative $(s'_1, \ldots, s'_L)$ for each light wavelength value. They compute the derivatives over the spectra normalized to unit area and improved by a Gaussian filter (Equation 9).

$$\mathrm{PD}\,(s,t) = \sum_{l=1}^{L} |s'_l - t'_l| \tag{9}$$

This measure has the same problem related to normalization discussed earlier. Moreover, PD ignores the information of spectral concavity to compare spectral shapes.

As we have shown in this section, there is no best spectral dissimilarity measure for all type of spectra. A good spectral dissimilarity for certain problems might be a bad measure in a different domain. In substances identification, experts visually compares UVS taking into account the similarity between spectral shapes. Based on this, we propose a measure that captures spectral shape with the aim of obtaining better results in spectra comparisons.

3 New Dissimilarity Measures for UVS Identification

The basic idea of Derivatives Sign Differences (DSD) is to count the points where either the monotony or the concavity of the spectra differs. Therefore, the lower value returned, the lesser spectral difference.

3.1 Derivatives Sign Differences Measure

Given two spectra $s = (s_1, \ldots, s_L)$ and $t = (t_1, \ldots, t_L)$, DSD computes the dissimilarity between s and t as follows:

1. Compute and smooth the first derivative value tuples from spectra s and t: $s' = (s'_1, \ldots, s'_L)$, $t' = (t'_1, \ldots, t'_L)$
2. Compute and smooth the second derivative value tuples from s' and t': $s'' = (s''_1, \ldots, s''_L)$, $t'' = (t''_1, \ldots, t''_L)$
3. Set $count = 0$
4. For each $l = 1, \ldots, L$
   - If $(sign\,(s'_l) \neq sign\,(t'_l) \vee sign\,(s''_l) \neq sign\,(t''_l))$ then $count = count + 1$
5. Return $count/L$

The normalization at step 5 returns a value in the [0, 1] interval. A median filter smoothes the derivatives tuples, using a window width of five values. We choose the window width experimentally, but small variations of this value attain similar results.

Notice in Fig. 5 (a) the decreasing monotony of a spectrum in light wavelength interval [280, 320]. Fig. 5 (b) shows a zoom of the interval [305, 310] where dark dots represent the spectrum points, and line segments represent the slope directions at each point. Opposite to what we would expect, a zoom of this interval shows heterogeneous slope directions. That is why we apply a median filter over the derivative values tuples (steps 1 and 2), achieving homogeneous

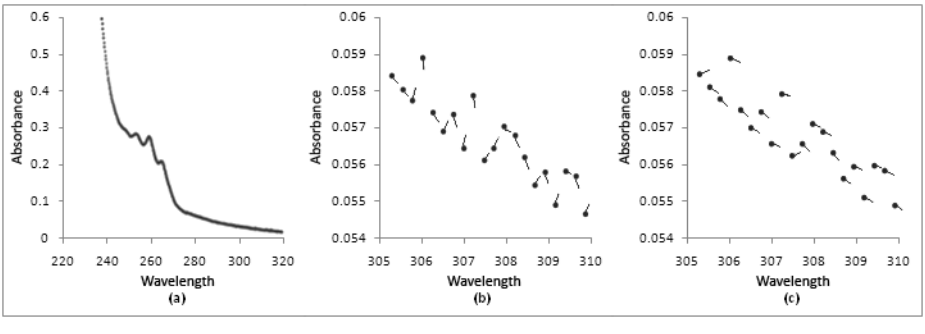

Fig. 5. Result of applying a median filter over the first derivatives values tuple. (a) Original spectrum. (b) Slope directions in a zoomed portion of the spectrum. (c) Homogeneous directions after applying a median filter.

slope directions, as shown in Fig. 5 (c). As we can see in the experimental section, this procedure improves the accuracy of the results.

Due to spectra derivatives are pre calculated and stored before searching for a query spectrum, the time complexity of DSD is $O(n)$, being n the total wavelength values. DSD effectively matches spectra from the same substance at different concentration. The proposed measure does not compare absorbance values but the signs of first and second derivatives tuples. Thus, DSD correctly matches different spectra from the same substance because these spectra do not differ in monotony and concavity.

In the experimentation, the proposed measure DSD outperforms all other dissimilarities except SCM. As SCM does not clearly outperform DSD, we decided to combine them to improve their results.

3.2 Fusion of DSD and SCM Dissimilarities

DSD and SCM compare spectra in different manners. DSD takes into account monotony and concavity of spectra, while SCM determines the correlations between the spectra tuples. We decided to combine them in order to benefit from their advantages and reducing their limitations. To do this, we use a combination scheme that returns small spectra difference if one of these measures indicates it, regardless of the result of the other measure.

Multiplying the results of two dissimilarity measures, both defined on the [0, 1] interval, if one of them returns a value close to zero, it takes precedence over the other; the final value will also be close to zero, indicating small difference. That is why we use the product rule [15] to combine DSD with SCM. This rule is one of the most used schemas for comparison functions combination.

$$\mathrm{DSD} - \mathrm{SCM}\,(s,t) = \mathrm{DSD}\,(s,t) \cdot \mathrm{SCM}\,(s,t) \tag{10}$$

Using the product rule we attempt to favor the best result of both dissimilarities achieving lower values when both dissimilarities agree in a good result.

4 Experimental Results

We tested the proposed dissimilarity measure on a database containing 206 ultraviolet spectra from 103 substances. For each substance, we extracted two spectra at different concentrations (Fig. 2) with the Spectrometer Cintra 101 [6]. The wavelength values ranges from 200 nm to 350 nm. The measurement interval was 0.5 nm to emphasize spectra details, and the average speed was 500 nm/min.

As our goal is to identify a query spectrum with its corresponding substance, for each query spectrum we sort all the spectra in the database in increasing order according to their dissimilarity with the query. A good measure always returns the correct substance in the first positions. That is why, in order to evaluate the performance of our dissimilarity, we build a ROC curve using a leave one out [16] sampling.

A ROC curve [13] is a useful tool in the evaluation of comparison functions for objects identification, this curve measures the tradeoff between correct and false identification rates. Moreover, it has become a standard in areas such as fingerprint identification [17], which is analogous to substance identification.

In terms of ROC curve, a dissimilarity measure is better than another one if it has higher true positive rates for most false positive rates.

Our first experiment shows how DSD with smoothed first and second derivatives outperforms the same dissimilarity without smoothing (Fig. 6).

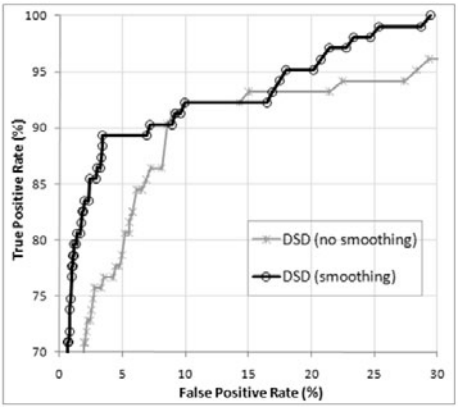

Fig. 6. The ROC curves of DSD smoothing derivatives versus DSD without smoothing

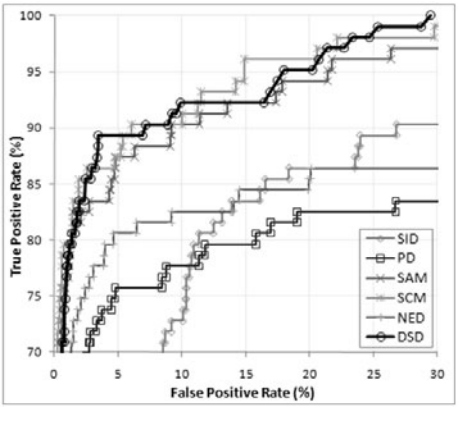

Fig. 7. ROC curves of DSD and all single dissimilarities

We compare DSD against the single dissimilarities reviewed in section 2, see Fig. 7. Notice that DSD clearly outperforms every other dissimilarity except SCM. However, SCM does not clearly outperform DSD.

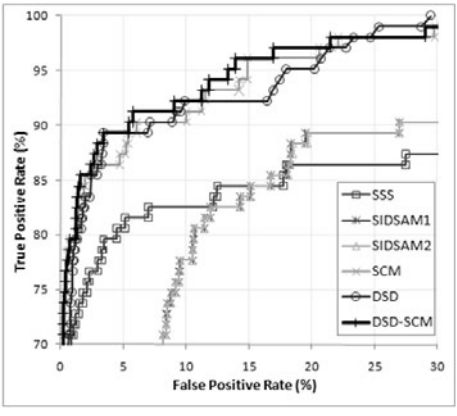

Fig. 8. ROC curves of combining DSD and SCM and three compound UVS dissimilarities

Our final experiment shows how the combination of DSD with SCM is superior to all individual measures and other composite measures proposed in the literature, as shown in Fig. 8.

The combination of DSD with SCM effectively takes the advantage of both measures and reduces its limitations. It returns low dissimilarity values if two spectra are dissimilar according to DSD or SCM. In addition, notice that combinations reviewed do not outperform its component measures.

5 Conclusions

In this paper, we introduce a new dissimilarity measure to compare UVS spectra. We create DSD based in the way experts visually compare spectra shape for substance identification. DSD allows better discrimination between UVS from different substances than most of the measures reviewed do. The combination of DSD with SCM clearly outperforms all single and combined measures analyzed in this paper. As future work, we plan to investigate our approach in similar problems using other wavelengths like visible and near infrared.

References

1. Paclík, P., Duin, R.P.W.: Classifying spectral data using relational representation. In: International Workshop on Spectral Imaging, pp. 31–34 (2003)
2. Demir, B., Ertürk, S.: Improved classification and segmentation of hyperspectral images using spectral warping. Int. J. Remote Sens. 29, 3657–3663 (2008)
3. Paclík, P., Leitner, R., Duin, R.P.W.: A study on design of object sorting algorithms in the industrial application using hyperspectral imaging. J. Real-Time Image Proc. 1, 101–108 (2006)

4. Van der Meer, F.: The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery. Int. J. Appl. Earth Obs. Geoinf. 8, 3–17 (2006)
5. Ingle, J.D.J., Crouch, S.R.: Spectrochemical Analysis. Prentice Hall, New Jersey (1988)
6. GBC UV-Visible Cintra 101/202/303/404 Spectrometer Operation Manual. GBC Scientific Equipment Pty Ltd., Australia, GBC part number: 01-0831-01 (2005)
7. Kruse, F.A., Lefkoff, A.B., Boardman, A.B., Heidebrecht, K.B., Shapiro, A.T., Barloon, P.J., Goetz, A.F.H.: The Spectral Image Processing System (SIPS) - interactive visualization and analysis of imaging spectrometer data. Remote Sens. Environ. 44, 145–163 (1993)
8. Van der Meer, F., Bakker, W.: Cross correlogram spectral matching (CCSM): application to surface mineralogical mapping using AVIRIS data from Cuprite, Nevada. Remote Sensing Environ. 61(3), 371–382 (1997)
9. Chang, C.-I.: An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis. IEEE Trans. Inf. Theory 46, 1927–1932 (2000)
10. Yingzi, D., Chein, I.C., Hsuan, R., Chein-Chi, C., James, O.J., Francis, M.D.A.: New hyperspectral discrimination measure for spectral characterization. Opt. Eng. 43, 1777–1786 (2004)
11. Homayouni, S., Michel, R.: Hyperspectral image analysis for material mapping using spectral matching. In: Proceedings of the XX ISPRS Congress, Proceedings volume IAPRS, Istanbul, July 12-23, vol. XXXV (2004)
12. Robila, S.A., Gershman, A.: Spectral matching accuracy in processing hyperspectral data. In: International Symposium on Signals, Circuits and Systems (ISSCS 2005), pp. 163–166 (2005)
13. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874 (2006)
14. de Carvalho, O.A., Meneses, P.R.: Spectral Correlation Mapper (SCM): An Improvement on the Spectral Angle Mapper (SAM). In: NASA JPL AVIRIS Wkshp. (2000)
15. Kuncheva, L.I.: Combining pattern classifiers: methods and algorithms. Wiley Interscience, Hoboken (2004)
16. Devroye, L., Gyorfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, New York (1996)
17. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition, 2nd edn. Springer, London (2009)