# A Spectrum Matching Algorithm Based on Dynamic Spectral Distance

Qi Jia, ShouHong Cao

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China

jiaqi199167@163.com, caoshouhong@foxmail.com

**Keywords:** Spectrum Matching, Hamming Distance, Difference of Spectrum, Similarity Measure

**Abstract .**Spectral matching algorithms are widely used in various models for spectral analysis and identification. Traditional spectral matching algorithms usually adopt the similarity of their absorbance value as the index to evaluate the similarity between two spectra. The absorbance value is related to concentration of substance and state at the sampling time, hence those algorithms have higher requirements on samples. This thesis proposes an algorithm based on dynamic spectral distance (DSD), considering both the waveform similarity and the absolute difference, taking the segmented hamming distance of derivative spectrum as the waveform difference measure between two spectral curves and the standard deviation of Difference of Spectrum (DS) as the absolute difference measure of spectrum. As the experiment shows, the algorithm can effectively identify not only spectral with similar waveform but also those with different waveform, and thus has higher accuracy and stability than traditional algorithms. The algorithm can be used in classification and matching system for UV-visible spectrum, infrared spectrum, hyper-spectrum, etc.

## Introduction

Spectral matching algorithms are widely used in various spectral analysis models and spectral database system. According to their principle, current matching algorithms can be divided into geometric measure method, probability measure method, transformation measure method[3], and the former two methods have higher precision requirements on absorbance or reflectance value. Due to the influence of noise, environment, state as well as other factors, even spectra acquired from the same substance may also be different in the process of spectral acquisition. Therefore when there are a large number of sample sets, spectral classification based on the above two methods tend to be inaccurate. In recent years, many researchers study the spectrum matching algorithm based on transformation measure[1], which is more stable than algorithms based on geometric measure or probability measure.Zhou Wanhuai from Zhejiang University put forward a full spectral matching algorithm based on Jaccard similarity coefficient[1]. This algorithm uses main factors to calculate Jaccard similarity coefficient between two spectra, however, it has some ambiguity and poor identification on spectra with narrow peak or overlapped peaks. Zhang Qilin, Wang Xianpei et al. put forward an infrared spectrum matching algorithm based on similarity system theory[2]. This algorithm takes the elasticity of peak number as similarity measure to obtain the relative difference between two spectra. Liu Wei et al. put forward the concept of DS[3], taking its information entropy and canberra distance as similarity measure.

On the basis of former researches, this thesis proposes a matching algorithm based on DSD by use of derivative spectrum and DS. In this algorithm, the waveform of spectral curves is taken as the index to measure their waveform similarity and the standard deviation of DS as the index to measure their absolute difference, and then the dynamic distance between spectra is calculated by

 

the above two indexes according to a certain proportion. In addition, the effectiveness of the algorithm is demonstrated by infrared spectrum of several common substances.

**The Matching Algorithm based on DSD**

In the spectrum matching algorithm based on DSD, the absolute difference of spectral curve on the vertical axis is related with many factors, so identification error will be caused if the difference on the vertical axis is used to calculate the distance between spectra. Thus, this thesis employs the difference of their increase/decrease characteristics to measure the distance between spectra. If within an area two spectra has greater difference in terms of their increase/decrease characteristics, we believe there's a big difference between the two spectra; and if their increase/decrease characteristics in an area are basically the same, we believe their difference is smaller and might be caused by concentration of substance.

In the design of spectrum matching algorithm, the absolute distance of spectral curves and differences of waveform should be fully considered. When there are significant differences in their waveform, it means there are also great differences in the internal structures of the two kinds of substances. When their characteristic peak position and waveform are similar, we believe the two substances belong to one kind but different breeds. The following presents the matching algorithm based on DSD including two stages, i.e. pre-processing and spectrum matching.

**Pre-processing.**The plus or minus characteristic of the first derivative determines the increase/decrease characteristics of spectral curves. To calculate the difference between two spectral curves, firstly conduct binarization processing to the first derivative. Take the first derivative of (n-1) points, then process it according to the following formula:

$$F_i = \begin{cases} 1 & k_i > 0 \\ 0 & k_i > 0 \\ F_{i-1} & k_i = 0 \end{cases} \quad (1)$$

Thereinto, $k_i$ is the first derivative at $i$ point. When $k_i = 0$, the slope of the tangent is 0 and extreme point might appear. For ease of calculation, the point where $k_i = 0$ is regarded as continuation of increase/decrease characteristics of the previous segment of spectral curve, so let's make it $F_{i-1}$.

After completion of binary-coding, segment $F_i$. For spectral curves, it's the position and number of characteristic peak that determine the structure of substance, so segment $F_i$ according to the characteristic peak. Besides, segment standard spectrum according to the segments of spectrum to be measured so that the number of segments on two curves and data points within each segment are equal.

**Dynamic Spectral Distance.**Hamming distance is a concept in the information theory, expressing the number of different characters at the corresponding positions between two strings with equal length. Calculate the hamming distance between spectrum to be measured and standard

spectrum in the i segment with the following formula:

$$d_i = \sum F_{1i} \oplus F_{2i} \quad (2)$$

Thereinto, $F_{1i}$、$F_{2i}$ are the first derivative of spectrum to be measured and standard spectrum after binarization processing at the i point within their respective segment. $\oplus$ is the XOR symbol.

The hamming distance between two spectral curves in the i segment can be obtained:

$$F_i = \frac{d_i}{l} * 100 \quad (3)$$

Thereinto,l is the number of data points within current segment. Multiplication by 100 is to quantify the value between 0 and 100.

Liu Wei et al put forward the concept of DS[3], which is a curve obtained by subtracting ordinate values of all the points in the two spectral curves one by one. If spectrum to be measured and standard spectrum have similar forms and peak positions, DS should be a straight line with a slope of nearly 0. And when DS is a curve with significant fluctuations, there should also be great difference in their waveform. Standard deviation of DS is used to measure the dispersion degree of spectral curve. The formula is:

$$\sigma_i = \frac{\sqrt{\Sigma(y_{ij} - \overline{y_l})^2}}{l} \quad (4)$$

Thereinto, $y_{ij}$ is the absorbance value of each point in each segment, $\overline{y_l}$ is the average absorbance value in each segment, and l is the number of data point in each segment.

Dynamic spectral distance (DSD) is proposed based on the above discussion with the calculation formula as follows:

$$F = \sum \sqrt{a * {F_i}^2 + (1-a) * {\sigma_i}^2} \quad (0 < a < 1) \quad (5)$$

Thereinto, a is a factor that can be adjusted dynamically, $F_i$ mainly emphasizes the difference in waveform (its value will be bigger when there's a greater difference in waveform of the two

curves) and $\sigma_i$ mainly emphasizes the difference in value. Value of a can be adjusted dynamically as required. For example when calculating the hamming distance between two spectral curves in advance, we can properly increase its value if the two spectral curves are quite different in their waveform and reduce its value if their waveform is similar and in this condition, spectral curves have similar waveform and their difference is mainly embodied in dispersion degree of DS. This algorithm considers difference in both spectral waveform and absorbance, and at the same time calculates the difference according to each characteristic peak, effectively reflecting the whole difference of spectrum. The following experiment applies the algorithm to validate its correctness.

**Experiment**

Near infrared spectrum acquisition and IRTracer-100 Fourier transform infrared spectrometer. After KBr squash technique is adopted to make squash slide, scan each squash slide for 30 times and scan background spectrum once for every 60 minutes so as to eliminate baseline drift.

Standard spectrum adopts: 3 different breeds of potato namely Hubei No.1 Potato produced in Enshi, Netherlands 15 produced in Anqiu of Shandong, Qingshu 168 produced in Dingxi of Gansu, and 3 kinds of drugs namely aspirin zinc, etofesalamide and acemetacin. Potatoes are purchased from the market and drugs are essential medicines in laboratory. After squashed by KBr technique, use infrared spectrometer to acquire spectrum. Two groups are divided for test respectively with classes of IR1-IR3 and classed of IR4-IR6,5 samples in each class are taken as standard spectrum and 15 samples as sets to be measured.

Smooth the moving window for spectrum to be measured so as to eliminate noisy data. Respectively make a 0.1~0.9 and verify the accuracy of IR1-IR3 potatoes and IR4-IR5 drugs. Their accuracy rates are shown in the Fig 1 and Fig 2:

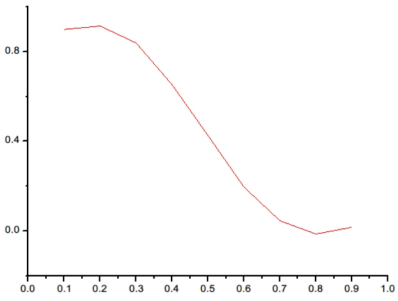


Fig 1.3 breeds of potatoes

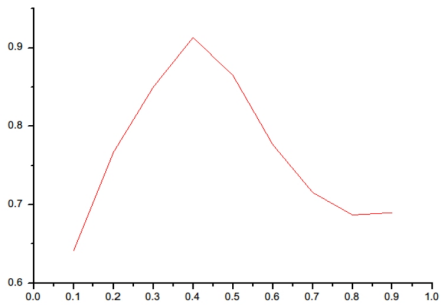


Fig 2. 3 kinds of drugs

As can be seen, for the IR1-IR3 potatoes, when $a = 0.2$, the accuracy of DSD algorithm is maximized. Since three potatoes belong to the same substance, the peak position of its spectral curve is basically the same and the main difference is embodied in the form of characteristic peak. When the value of a is smaller, DSD algorithm mainly considers the absolute difference of spectral curve. For the IR4-IR6 drugs, when $a = 0.4$, the accuracy of DSD algorithm is maximized. Thus it can be seen, better classification results can be obtained if peak form and absolute difference of spectrum are comprehensively considered.

Table 1 compares the DSD algorithm with conventional ED(Euclidean Distance) algorithm and CC(Correlation Coefficient) algorithm.

Accuracy of ED、CC and DSD

| | ED | CC | DSD |
|---|---|---|---|
| IR1、IR4 | 72.30% | 73.70% | 93.20% |
| IR1、IR2 | 69.50% | 72.40% | 86.40% |
| IR1、IR6 | 75.40% | 73.10% | 88.50% |
| IR2、IR4 | 73.80% | 75.30% | 90.50% |
| IR3、IR5 | 75.10% | 71.00% | 86.40% |
| IR1、IR4、IR5 | 65.20% | 70.50% | 84.70% |
| IR1、IR4、IR6 | 63.40% | 71.60% | 85.40% |
| IR2、IR4、IR5 | 63.60% | 68.20% | 84.10% |
| IR3、IR4、IR5 | 61.10% | 67.50% | 85.00% |
| IR1、IR4、IR5、IR6 | 59.80% | 67.30% | 83.70% |
| IR1、IR2、IR4、IR5 | 6.50% | 68.60% | 84.50% |
| IR2、IR3、IR4、IR5 | 63.70% | 65.30% | 80.90% |
| IR1、IR2、IR4、IR5、IR6 | 59.30% | 62.10% | 79.20% |

Table 1

In the classification and matching of near-infrared spectrum of samples with ED algorithm, CC algorithm and DSD algorithm, the relationship between their accuracy rates and the number of classes are shown in the following Fig 3:

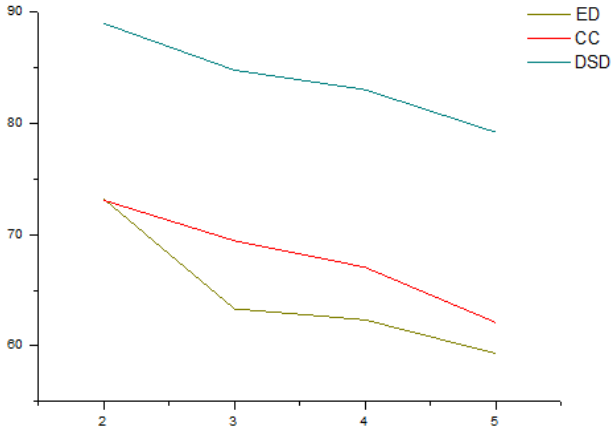


Fig 3. Comparison of DSD, ED, and CC

The accuracy rate of DSD algorithm is much higher than that of ED algorithm and CC algorithm. With the number of classes increasing, the accuracy rate of ED algorithm and CC algorithm becomes lower while DSD algorithm still ensures a high accuracy. Nevertheless, the accuracy of DSD algorithm also shows a downward trend.

After research and analysis, we find that due to the wide band of near-infrared spectrum and lots of information contained in the spectrum, weighted spectrum often has a big hamming distance yet DS has a small value in calculating the DSD of near-infrared spectrum, so even if dynamic factors can be adjusted, the error cannot be eliminated and correct classification cannot be obtained. Principal component analysis (PCA) can be introduced based on DSD algorithm. Reduce the near-infrared spectrum data dimension, and then calculate the weighted hamming distance so as to lower the influence of some extraneous data on the results.

In general, when the number of classes is small, DSD algorithm shows good performance in infrared spectrum classification and meets the requirement. With the number of classes continuously increasing, the accuracy rate of DSD algorithm is also declining, but still maintains a higher accuracy. The DSD algorithm meets the demands of some classification models when classification accuracy is not strictly limited.

## Conclusions

This thesis proposes the DSD algorithm. Unlike traditional algorithms which have higher requirements on spectral absorbance (or reflectance value), this algorithm considers both the waveform similarity and the absolute difference of spectral curves. The generalized distance between two spectra is calculated by the above two indexes according to a certain proportion, which can be dynamically adjusted according to concrete requirements and experiment. As the experiment shows, the DSD algorithm could effectively reflect the waveform difference of spectral curves and provide reference for design of spectral database system and online matching system.

## Acknowledgements

This work is supported by NSFC (Grant Nos. 61300181, 61502044), the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

## References

[1] Zhou Wanhuai. Key algorithms for apple near-infrared spectral database and development of the prototype system [D]. Zhejiang University, 2014.

[2] Zhang Qilin, Wang Xianpei, Zhao Yu et al.Spectrogram comparison method for infrared spectrum based on similarity system theory [J]. Chinese Journal of Spectroscopy Laboratory, 2013, 30(6). DOI:10.3969/j.issn. 1004-8138. 2013.06. 006.

[3] Liu Wei, Xu Shuiping, Yuanjun et al. Improved method for spectral similarity measure based on information entropy of Difference of Spectrum[C]//Papers selection from the 2nd frontier technology BBS of geomatics science and technology. 2008.

[4] Wen Binggong, Feng Wufa, Liu Wei et al. Matching and Classification Based on the Whole Comparability Measure of Spectral Curve[J]. Journal of Geomatics Science and Technology, 2009, 26(2):128-131. DOI:10.3969/j.issn. 1673-6338. 2009. 02.014.

[5] Zhou W, Ying Y, Xie L. Spectral Database Systems: A Review[J]. Applied Spectroscopy Reviews, 2012, 47(8):654-670.

[6] Karpushkin E, Bogomolov A. New system for computer-aided infrared and Raman spectrum interpretation[J]. Chemometrics & Intelligent Laboratory Systems, 2007, 88(1):107-117.

[7] Fu Jing, Shu Ning, Kong Xiang Bing. Spectrum matching algorithm based on hidden Markov model[C]//Abstract books of the 17th conference on remote sensing in China. 2010.

[8] Zhang Xinle, Zhang Shuwen, Li Ying et al. Extracting Black Soil Border in Heilongjiang Province Based on Spectral Angle Match Method[J]. Spectroscopy and Spectral Analysis, 2009, 29(4):1056-1059. DOI:10.3964/j.issn. 1000-0593 (2009)04-1056-04.

[9] Luinge H J, Leussink E D, Visser T. Trace-level identity confirmation from infrared spectra by library searching and artificial neural networks[J]. Analytica Chimica Acta, 1997, 345(1):173–184.