электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК.554

Определение оптимального алгоритма для идентификации веществ по спектрам

А.А. Кудрявцев

Студент,кафедра «Физика» МГТУ им. Н.Э. Баумана, г.Москва, Россия

Научный руководитель: Морозов А.Н., д.ф.-м.н., проф. кафедры «Техническая физика» МГТУ им.Н.Э. Баумана

ЗАО «ЦПФ»

Vatutu@gmail.com

Введение

В настоящее время неразрешенным остался вопрос о безошибочной идентификации веществ по спектрам. Существует множество различных методик, которые сравнивают два спектра и на выходе получают численное значение, определяющее степень схожести двух функций. Проблема заключается в том, что при некоторых обстоятельствах алгоритмы идентификации выдают ошибочные значения. Таким образом, не давая точно определить, с каким веществом мы имеем дело. К таким обстоятельствам относятся различные спектры одинаковых веществ, но разной концентрации или сильно зашумленные спектры, а так же спектры различных веществ, которые обладают схожими спектральными формами. В данной работе рассмотрена эффективность различных методик на сильно зашумленных спектрах люминесценции и представлено действие этих алгоритмов на ультрафиолетовых спектрах без шумов. Показана разность между эффективностями одних и тех же методик на спектры разных диапазонов и шумов.

Теоретическая часть

Целью данной работы является анализ эффективности существующих методик идентификации веществ по двум спектрам. Имеются два спектра одного вещества, один из них эталонный, взят из базы данных, а другой - зашумленный. Необходимо узнать к

какому веществу принадлежат эти спектры. Для этого существуют специальные функции сравнения, которые сопоставляют два спектра и определяют их степень схожести. Ниже изложены алгоритмы для идентификации веществ по двум спектрам, на которых будет проводиться анализ их эффективности.

Для анализа эффективности методик имеем 2 набора спектров одинаковых веществ. Первый набор спектров представляет собой чистые вещества, второй — искусственно зашумленные, то есть отношение сигнал/шум хуже, чем у первого набора. Затем применяем функции идентификации к заданным наборам спектров, находим численные значения и заносим их в матрицу. По столбцам данной матрицы располагаются чистые спектры, а по строкам зашумленные. В этой матрице находятся все возможные коэффициенты схожести спектров друг с другом. На главную диагональ матрицы попадают два спектра одного и того же вещества, но один из них зашумлен. Таким образом, чем ближе элементы главной диагонали матрицы к нулю, а остальные элементы отличны от нуля, тем лучше работает функция. Дальше строится *ROC*-кривая, которая и является основным показателем эффективности методик.

ROC-кривая – это функция, которая показывает отношение верных срабатываний к ложным [1]. По оси ординат откладываются верные срабатывания, а по оси абсцисс – ложные срабатывания в процентом соотношении от нуля до 100 %. Главная диагональ, которая разбивает все пространство *ROC*-кривой на два треугольника (аналог функции $y = x$) соответствует случайному угадыванию вещества. То есть, как будто мы не сравниваем два спектра с помощью какого-либо алгоритма, а подбрасываем монету и определяем по «орлу» и «решке» совпадение и несовпадение веществ. Имеют места только те функции, которые расположены в верхнем треугольнике, чтобы превзойти случайное угадывание [2]. Если функция оказалась в нижнем треугольнике, что соответствует результату хуже, чем случайному, то либо эта методика не имеет смысла либо ошибка в алгоритме построения кривой, и следует симметрично отобразить ее относительно главной диагонали. Главным показателем эффективности методик обнаружения веществ по спектрам является площадь под кривой (под *ROC*-кривой). Чем она больше, тем лучше алгоритм.

Краткий обзор используемых алгоритмов

Представлены функции сравнения, которые были исследованы для выявления наиболее эффективного метода обнаружения веществ. Каждый из приведенных алгоритмов зависит от двух переменных s и t. Они представляют собой значения интенсивности двух спектров, представленных как набор чисел $s = (s_1, \dots, s_l)$ и $t =$

$(t_1, \dots, t_l)$ одинаковой длинны. L – количество компонентов каждого из спектров. Подробное описание каждой из функции можно найти в работе [3].

1. ***Спектральное угловое отклонение - Spectral Angle Mapper (SAM)***

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

2. ***Спектральное корреляционное измерение - Spectral Correlation Measure (SCM)***

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

3. ***Спектральная информационная дивергенция - Spectral Information Divergence (SID)***

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

Где $p_k = s_k / \sum_{l=1}^{L} s_l$ и $q_k = t_k / \sum_{l=1}^{L} t_l$.

4. ***Спектральная схожесть пространства - Spectral Similarity Scale (SSS)***

$$SSS(s,t) = \sqrt{\frac{1}{L\sum_{l=1}^{L}(s_l - t_l)^2} + (1-r)^2}$$

Где

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

5. ***Нормализованное евклидово пространство - Normalized Euclidean Distance (NED)***

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

6. *Производные различных знаков - Derivatives Sign Differences (DSD)*

Алгоритм:

1) Вычисляем и сглаживаем первые производные двух спектров
2) Вычисляем и сглаживаем вторые производные от спектров
3) Присваиваем $count = 0$
4) Для каждого $l = 1, \dots, L$
   - Если $\left(sign(s_l') \neq sign(t_l')$ и $sign(s_l'') \neq sign(t_l'')\right)$, то $count = count + 1$
5) Возвращаем $count/L$

Результаты

В работе [3] кроме описания функций так же проводится анализ эффективности методик. Они использовали чистые ультрафиолетовые спектры поглощения без шумов.

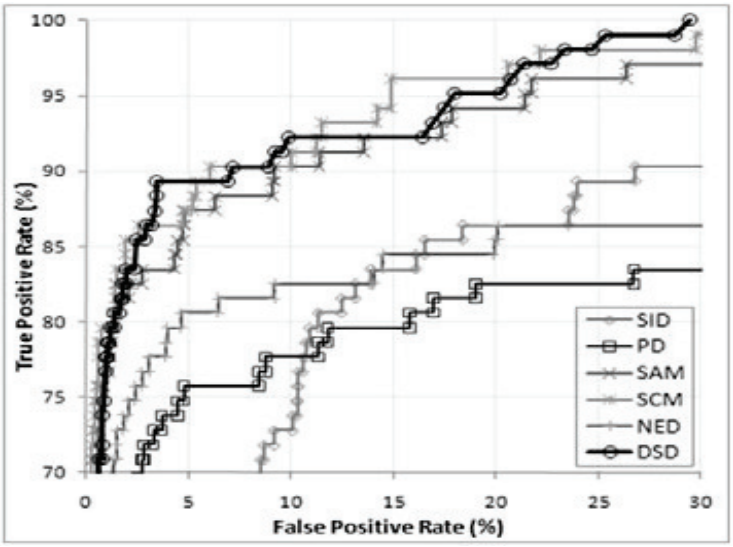

Рис. 1. *ROC*-кривые из работы [3]

На рис. 1 показана эффективность методик на ультрафиолетовых спектрах. На нем видно, что функция *DSD* является одним из наиболее эффективных алгоритмов сравнения, так как располагается выше других функций (имеет большую площадь). Именно на этом методе акцентируется внимание в той работе, как на наилучшем методе для сравнения ультрафиолетовых спектров. Так же хорошее отношение верных срабатываний к ложным

имеют алгоритмы *SCM* и *SAM*. Хуже всего показали себя такие функции как *SID* и *NED*, так как их кривые расположены ниже других и соответственно имеют меньшую площадь.

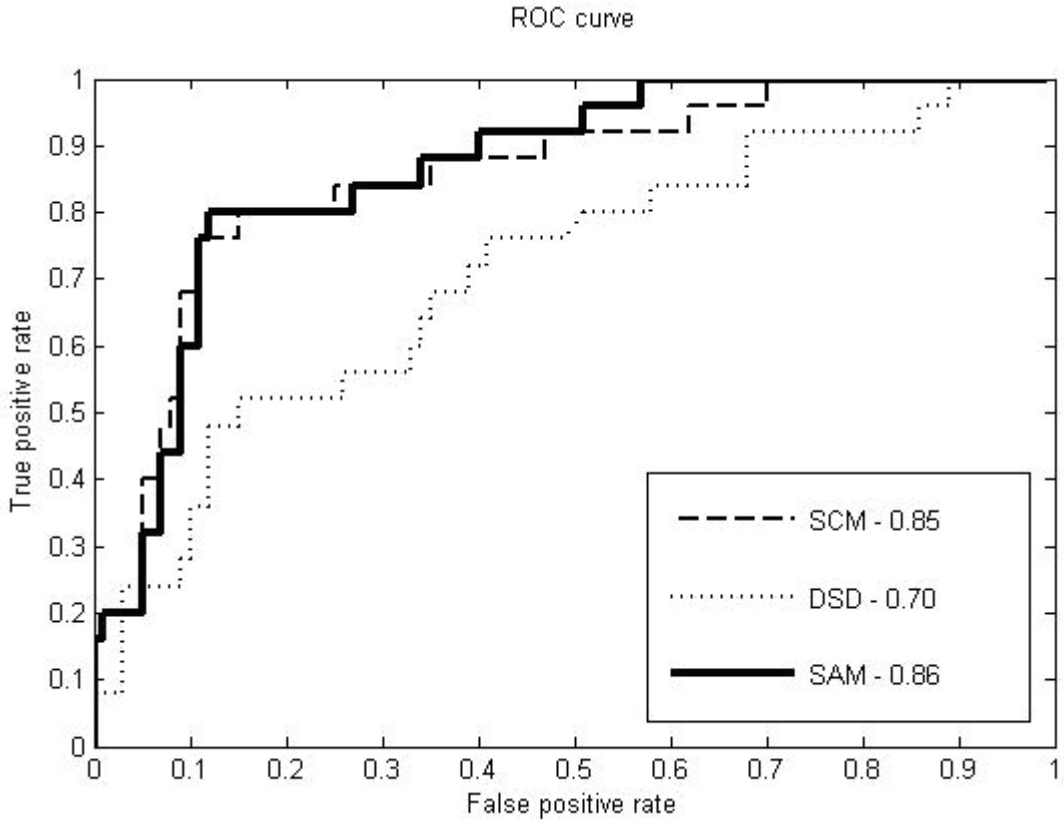

Рис. 2. *ROC* – кривые для алгоритмов, использующих различные меры несоответствия

На рис. 2 показана эффективность рассмотренных алгоритмов на сильно зашумленных спектрах люминесценции. Справа в нижнем углу представлено численное значение площади под каждой из кривой. Видно, что функции *SAM* и *SCM* имеют большую площадь, чем алгоритм *DSD*, соответственно более эффективны. *DSD* как на ранних так и на поздних стадиях имеет большую вероятность ошибки, чем другие две представленные функции. Так же кривые двух функций практически совпадают и обладают схожей площадью, таким образом, нельзя однозначно сказать какой из этих алгоритмов лучше. Для полного анализа ниже представлены методики, не включенные в рис. 2 для визуального удобства.

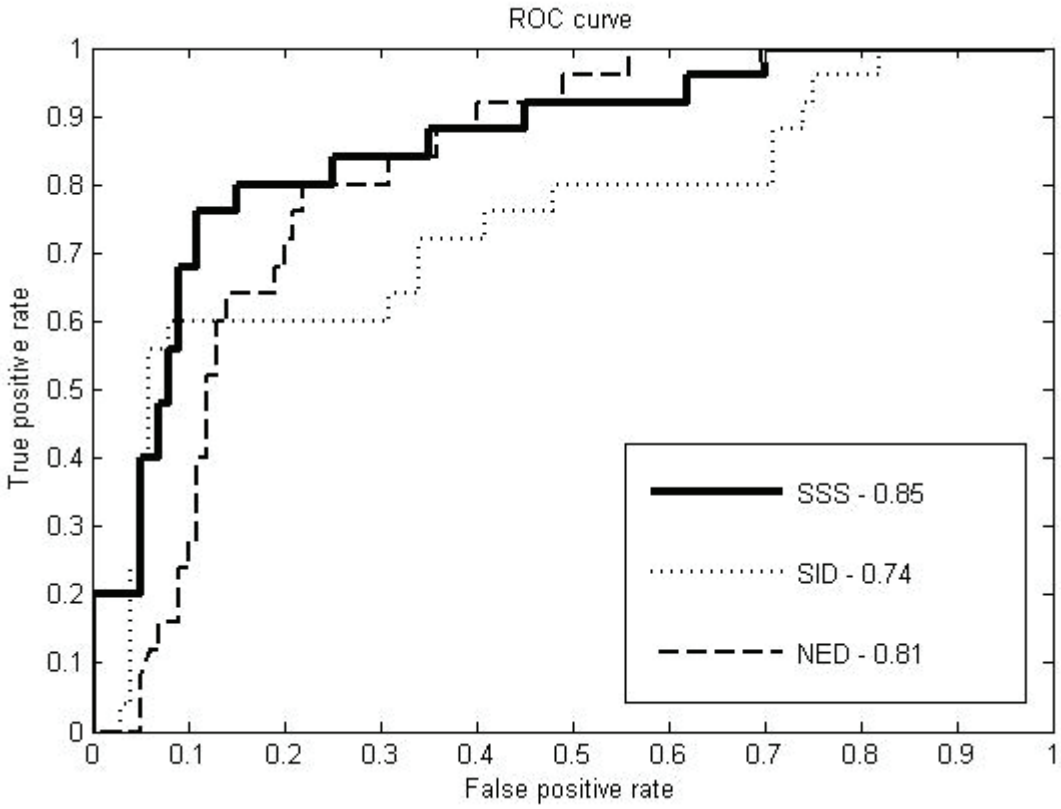

Рис. 3. *ROC* – кривые для алгоритмов, использующих различные меры несоответствия

На рис. 3 показана эффективность таких методик как *SSS, SID, NED*. Все три метода по эффективности превосходят *DSD*. Абсолютно лучшей среди данных трех алгоритмов является *SSS*, имеющая большую площадь под кривой. Остальные функции немного хуже, но стоит отметить, что кривая *NED* достигает 100 % верных срабатываний немного раньше, чем SSS, но на ранней стадии дает ошибочные результаты. *SID* достигнув 60 % вероятности верных срабатываний в дальнейшем выдает ошибочные результаты и достигает 100 % верных срабатываний при 82 % ложных.

Из рис. 2 и рис. 3 видно, что наибольшую площадь под *ROC*-кривой имеют – *SCM, SAM, SSS*, что означает их наибольшую эффективность.

Заключение

Был проведен анализ эффективности различных алгоритмов идентификации веществ на спектрах с различной степенью зашумления. Рассмотрены следующие меры несоответствия: *SCM, SAM, DSD, SSS, SID, NED*. В одном случае использовались чистые ультрафиолетовые спектры поглощения со слабыми шумами, а в другом – сильно зашумленные спектры люминесценции. Основное отличие, конечно же, заключается не в

диапазоне спектров, а в отношении сигнал/шум. Показано, что функция (*DSD*), имеющая высокую селективную способность на чистых спектрах, имеет слабую эффективность на сильно зашумленных спектрах. На сильно зашумленных спектрах наиболее эффективными оказались алгоритмы *SAM, SCM, SSS*.

Список литературы

1. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874 (2006)
2. Devroye, L., Gyorfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, New York (1996)
3. Gutierrez-Rodriguez1 A.E., Medina-Perez M.A., Martinez-Trinidad J.F., Carrasco-Ochoa J.A., Garcia-Borroto M.: New Dissimilarity Measures for Ultraviolet Spectra Identification. Springer, Berlin(2010)
4. Статический фурье-спектрометр видимого диапазона Бойко А.Ю., Голяк И.С., Голяк И.С., Дворук С.К., Доровских А.М., Есаков А.А., Корниенко В.Н., Косенко Д.В., Кочиков И.В., Морозов А.Н., Светличный С.И., Табалин С.Е. Известия Российской академии наук. Энергетика. 2010. в"- 2. С. 12-21.
5. Статический фурье-спектрометр видимого и ближнего ультрафиолетового диапазонов спектра Бойко А.Ю., Голяк И.С., Голяк И.С., Дворук С.К., Доровских А.М., Есаков А.А., Корниенко В.Н., Косенко Д.В., Кочиков И.В., Морозов А.Н., Светличный С.И., Табалин С.Е. Вестник Московского государственного технического университета им. Н.Э. Баумана. Серия: Естественные науки. 2009. в"- 3. С. 10-28.
6. Методика получения и обработки спектральной информации с помощью статического фурье-спектрометра Глаголев К.В., Голяк И.С., Голяк И.С., Есаков А.А., Корниенко В.Н., Кочиков И.В., Морозов А.Н., Светличный С.И., Табалин С.Е. Оптика и спектроскопия. 2011. Т. 110. в"- 3. С. 486-492.