Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра физики полупроводников

Методы машинного обучения для анализа данных спектроскопии комбинационного рассеяния света

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студента 2 курса 202 группы

направления 11.04.04 «Электроника и наноэлектроника»

факультета нано- и биомедицинских технологий

Рябова Евгения Александровича

Научный руководитель
Доцент кафедры инноватики на базе АО «НЕФТЕМАШ»-САПКОН, к.ф.-м.н. Д.Н. Браташов

должность, уч. степень, уч. звание | подпись, дата | инициалы, фамилия

Зав. кафедрой
профессор, д.ф.-м.н. А.И. Михайлов

должность, уч. степень, уч. звание | подпись, дата | инициалы, фамилия

Саратов 2017

ВВЕДЕНИИ

Спектроскопия комбинационного рассеяния (КР) - один из важнейших методов молекулярной спектроскопии, в основе которого лежит явление комбинационного рассеяния света [1]. Этот метод имеет ряд преимуществ по сравнению с другими методами. Он является неразрушающим, отличается простой пробоподготовкой, не требует специальных условий для детектирования спектра.

В результате анализа изображения микроскопии КР получается массив из большого числа спектров. В этих спектрах могут присутствовать шумы, разные значения интенсивности, постоянный фон, случайные выбросы и т.д.. Перед анализом спектров необходимо устранить артефакты их получения. В результате должны получиться чистые спектры. При этом остается еще одна проблема: большое количество спектров. Для решения этой проблемы можно объединить схожие спектры в группу (кластер) и в результате для каждой группы находиться средний спектр. Эту задачу можно решить с помощью кластеризации (машинного обучения без учителя).

После этого можно распознавать спектры. Для распознавания спектра необходимо отнести его к одному из классов, присутствующих в базе данных. Это может быть сделано одним из методов классификации (машинного обучения с учителем) на основе обучения на примерах.

Формулируется **цель работы**, которая состоит в разработке системы хранения спектров комбинационного рассеяния света с поиском похожих спектров.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Провести литературный обзор по методам спектроскопии комбинационного рассеяния, методов классификации и кластеризации данных (машинного обучения);

2. Сформулировать математическое описание критериев сравнения различных спектров КР и алгоритмов их преобразования;

3. Создать базу данных спектров КР;

4. Выбрать алгоритм классификации для распознавания спектров и подобрать его параметры;

5. Выбрать алгоритм кластеризации работы с данными микроскопии.

Положение, выносимое на защиту

Для распознавания спектров комбинационного рассеяния света с помощью метода машины опорных векторов наилучший результат дает использование линейного ядра, параметра допустимой ошибки 50, предварительного нормирования спектров и генерации векторов признаков на основе метода главных компонент с количеством компонент, равным 50.

Структура магистерской работы состоит из введения, теоретической части, практической части, заключения, списка литературы (26 источников).

Общая характеристика работы

Во **введении** обосновывается тема, объясняется актуальность данной тематики, формулируется цель работы, состоит в разработке системы хранения спектров комбинационного рассеяния света с поиском похожих спектров.

В **теоретической части** проводится литературный обзор по методам спектроскопии комбинационного рассеяния, методов классификации и кластеризации данных (машинного обучения). Рассматриваются 2 модели КР: классическая модель КР и квантовая модель КР. В классической модели КР в веществе индуцируется дипольный момент при прохождении световой

(электромагнитной) волны за счет смещения электронов в поле волны от положения равновесия. Относительная интенсивность стоксовой и антистоксовой составляющей спектра КР, предсказываемая как отношение $[(w - w_0) / (w + w_0)]^4$, не согласуется с экспериментальными данными. Для правильного определения отношения интенсивностей стоксовой и антистоксовой составляющей спектра КР рассмотрена квантовая модель КР, в которой процесс состоит из двух связанных между собой актов - поглощения первичного фотона с энергией $h\nu$, и испускания фотона с энергией $h\nu'$. В условиях теплового равновесия заселенность колебательных подуровней подчиняется распределению Больцмана, таким образом заселенность на более высоких уровнях уменьшается по экспоненциальному закону. Следовательно, на первом колебательном подуровне заселенность будет гораздо меньше, чем на нулевом, что приводит к гораздо меньшей интенсивности антистоксовых линий в спектре КР по сравнению с интенсивностью стоксовых линий.

Рассмотрены разновидности комбинационного рассеяния света: резонансное, вынужденное, гигантское.

Рассмотрены приборы для регистрации КР. В качестве источника монохроматического света рассмотрен принцип работы лазера. Для регистрации спектра в спектрометре рассмотрена ПЗС-матрица.

Перед анализом спектров КР нужно убрать паразитные элементы, для этого были рассмотрены алгоритмы предобработки спектров.

Спектр математически можно рассматривать как вектор в N-мерном пространстве. Размерность определяется количеством значений в диапазоне спектра. Каждая ось в N-мерном пространстве отвечает за один элемент в векторе.

Для уменьшения размерности векторов с минимальной потерей информации, то есть извлечения только полезной информации, применяют метод главных компонент (PCA, principal component analysis). Суть PCA состоит в вычислении ковариационной матрицы для уменьшения избыточности данных и максимизации дисперсии.

Для того чтобы сравнить два спектра между собой, нужно вычислить расстояние между ними. Для вычисления расстояния между двумя объектами применяются различные меры сходства (меры подобия), называемые также метриками или функциями расстояний. Существует множество функций расстояний, вот лишь основные из них:

1. Евклидово расстояние

Наиболее распространенная функция расстояния. Представляет собой геометрическим расстоянием в многомерном пространстве. Это расстояние находится следующим образом:

$$\rho(x, x') = \sqrt{\sum_{i}^{n} (x_i - x_i')^2}$$

2. Картирование по спектральному углу (SAM, Spectral Angle Mapper)

Для сравнения используется угол между векторами. Поскольку SAM выдает результат сравнения как синус угла между векторами, результат сравнения меняется от 0 до 1 [15]. SAM находится по формуле:

$$p(x, x') = cos^{-1}\left(\frac{\sum_i^n x_i x'_i}{\sqrt{(\sum_i^n x_i{}^2)(\sum_i^n x'_i{}^2)}}\right)$$

Для распознавания спектров был рассмотрен классификатор «машина опорных векторов» (support vector machine, SVM). Рассмотрен SVM для случая «неперекрывающихся» классов, для случая «перекрывающихся» классов, применение ядер и многоклассовый классификатор.

Для анализа данных микроскопии были рассмотрены алгоритмы k-means, k-medians, DBSCAN.

В **практической части** были представлены результаты работы распознавания спектров КР и кластеризация.

Использована база данных (БД) спектров КР минералов из проекта RRUFF [24]. Исходная БД проекта состоит из 5168 спектров КР.

Для обучения SVM исходная БД была разбита на две выборки. Первая выборка, которая служит для обучения SVM, состоящая из 3497 спектров КР, является обучающейся выборкой. Вторая часть БД служит для проверки точности распознавания спектров алгоритмом SVM, состоит из 1671 спектров КР и названа тестовой выборкой.

Для реализации SVM и PCA использовался язык программирования Python и библиотека Scikit-learn [25]. Был создан модуль распознавания спектров КР для пакета свободного программного обеспечения анализа данных микроскопии Gwyddion [26].

Для всех спектров КР в БД был выбран диапазон ν от 600 $см^{-1}$ до 1600 $см^{-1}$ с шагом 1 $см^{-1}$. В дальнейшем данные нормировались, и использовался PCA для уменьшения избыточности данных.

Для обучения SVM использовались параметры C = 50, использовался PCA с количеством компонент равным 50 и параметром whiten = {False, True}. Точность распознавания SVM при whiten = False равно 0,5376, а при whiten = True равно 0,55502.

Для пакета свободного программного обеспечения анализа данных микроскопии Gwyddion созданы модули кластеризации объемных данных используя алгоритмы K-means и K-medians. Модули K-means и K-medians написаны на Си с использованием библиотек Glib и Gwyddion [26].

Выбор K-means связан с простотой реализации и быстротой работы. Модуль K-medians подобен K-means, но средний спектр считается через медиану. Для проверки модулей был выбран образец таблетки цитрамона-П, который был исследован на микроскопе Renishaw Invia, при возбуждении лазером с длиной волны 785 нм и мощностью 5 мВт в режиме Streamline, 180x120 точек данных, что требовало порядка 1 часа для сбора данных.

Для данного образца изменение метрики не дало заметных различий в результате работы модулей k-means и k-medians.

ЗАКЛЮЧЕНИЕ

В работе описываются использование методов спектроскопии комбинационного рассеяния света и методов машинного обучения для анализа спектров комбинационного рассеяния.

В теоретической части был проведен анализ литературы, таким образом были рассмотрены:

1. классическая и квантовая модель КР;

2. регистрация спектров КР;

3. предобработка спектров КР;

4. классификация на основе модели – «машина опорных векторов»;

5. кластеризация – к-средних, к-срединных, DBSCAN.

В практической части были созданы программы для анализа спектров комбинационного рассеяния света применяя методы машинного обучения:

1. для ознакомления с работой алгоритма SVM была написана программа на языке программирования C++;

2. для ПО Gwyddion был написан модуль распознавания спектров КР с использованием SVM, с использованием скриптового языка Python и

библиотеки Scikit-learn;

3. была исследована работа модуля распознавания спектров КР с использованием базы данных спектров КР минералов из проекта RRUFF, которая была разбита на обучающую выборку и на тестовую выборку;

4. для ПО Gwyddion были написаны модули k-means и k-medians, с использованием языка программирования Си и библиотеки Glib;

5. разница между этими метриками не дала заметных различий в результате работы модулей.

СПИСОК ЛИТЕРАТУРЫ

1. Пентин, Ю.А. Основы молекулярной спектроскопии [Текст] / Ю. А. Пентин, Г. М. Курамшина. // М. : Мир ; БИНОМ. Лаборатория знаний, 2008. – 398 с.

2. Демтредер, В. Современная лазерная спектроскопия. Пер. С англ.: Учебное пособие [Текст] / В. Демтредер // Долгопрудный: Издательский Дом «Интеллект», 2014. – 1072 с.

3. Емельянов, В.И. Эффект гигантского комбинационного рассеяния света молекулами, адсорбированными на поверхности металла [Текст] / В.И. Емельянов, Н.И. Коротеев // Успехи физических наук, 1981. Т. 135. вып. 2. С. 345–361

4. ИНТЕГРА Спектра II [Электронный ресурс] / [Б. м. : б. и.], – Режим доступа: http://www.ntmdt-si.ru/afm-raman/ntegra-spectra-2 (дата обращения: 10.06.2017).

5. Ying-Sing Li, Raman spectroscopy in the analysis of food and pharmaceutical nanomaterials [Text] / Ying-Sing Li, Jeffrey S. Church // Journal of food and drug analysis. 2014. vol. 22. n. 1. P. 29-48.

6. Спектрометр высокого разрешения HR4000 [Электронный ресурс] / [Б. м. : б. и.], – Режим доступа: http://www.oemoptic.ru/c11x.php (дата обращения: 10.06.2017).

7. Гелий-неоновый лазер [Текст] / [Б. м. : б. и.], – Режим доступа: https://fis.wikireading.ru/6381 (дата обращения: 10.06.2017).

8. Херман, Й. Лазеры сверхкоротких световых импульсов: Пер. с нем. [Текст] / Херман. Й., Вильгельми Б. // М.: Мир, 1986, — 368 с.

9. Салех, Б. Оптика и фотоника. принципы и применения. Пер. С англ.: Учебное пособие [Текст] / Салех Б., Тейх М. // Долгопрудный: Издательский Дом «Интеллект», 2012. – 784 с.

10. Laser Diode Working Principle [Text] / [S. l. : s. n.], 2016. – Access mode: http://engineeringtutorial.com/laser-diode-working-principle/ (online; accessed: 10.06.2017).

11. Кратко о п/п лазерах [Текст] / [Б. м. : б. и.], 2011. – Режим доступа: http://nag.ru/articles/reviews/20748/kratko-o-p-p-lazerah.html (дата обращения: 10.06.2017).

12. ПЗС-матрица [Текст] / [Б. м. : б. и.] – Режим доступа: www.wikiwand.com/ru/ПЗС-матрица html (дата обращения: 10.06.2017).

13. Компьютерные методы ввода и обработки полей со сложной пространственной структурой векторов [Текст] / [Б. м. : б. и.], 2015. – Режим доступа: http://optics.sinp.msu.ru/prak/p25a/prac25a.htm (дата обращения: 10.06.2017).

14. Classification of inflammatory bowel diseases by means of Raman spectroscopic imaging of epithelium cells [Text] / Bielecki C. et al. // Journal of biomedical optics, 2012. vol. 17. n. 7.

15. Гиперспектральное дистанционное зондирование в геологическом картировании [Текст] / Под науч. ред. докт. техн. наук, проф. Г.Г. Райкунова.

// М. : ФИЗМАЛИТ, 2014. – 136 с.

16. Классификация данных методом опорных векторов [электронный ресурс] // [Б. м. : б. и.], 2015. – URL: http://habrahabr.ru/post/105220/ (дата обращения: 06.05.2017)

17. Comparison of classifier methods: a case study in handwritten digit recognition [Text] / L. Bottou et al. // Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on. IEEE, 1994. vol. 2. P. 77-82.

18. Kreßel, U. H. G. Pairwise classification and support vector machines [Text] / U. H. G. Kreßel // Advances in kernel methods. MIT Press, 1999. P. 255-268.

19. Platt, J. C. Large margin DAGs for multiclass classification [Text] / J. C. Platt, N. Cristianini, J. Shawe-Taylor // Proceedings of the 12th International Conference on Neural Information Processing Systems. MIT press, 1999. P. 547-553.

20. Кластеризация: алгоритмы k-means и c-means [электронный ресурс] // [Б. м. : б. и.], 2015. – URL: http://habrahabr.ru/post/67078/ (дата обращения: 06.05.2017)

21. Бериков, В. С. Современные тенденции в кластерном анализе [Текст] / В. С. Бериков, Г. С. Лбов // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. — 26 с.

22. Вятченин, Д. А. Нечеткие методы автоматической классификации [Текст] / Д. А. Вятченин // Минск: Технопринт, 2004. — 219 с.

23. Ester M. A density-based algorithm for discovering clusters in large spatial databases with noise [Text] / Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu // Kdd., 1996. vol. 96. n. 34. P. 226-231.

24. The power of databases: The RRUFF project [Text] / Lafuente B., Downs R.

T., Yang H., Stone N. // Highlights in Mineralogical Crystallography / Armbruster T. and Danisi R. M.. Berlin, Germany : W. De Gruyter. 2015 P. 1-30

25. Scikit-learn: Machine Learning in Python [Text] / Pedregosa et al. // JMLR 2011. 12, P. 2825-2830

26. David Nečas, Gwyddion: an open-source software for SPM data analysis [Text] / David Nečas, Petr Klapetek // Cent. Eur. J. Phys. 2012. vol.10. n. 1, P. 181-188