Генерация признаков из хромато-масс-спектрограмм при помощи кластеризации пиков для решения задач классификации в липидомике

Василюк А. В.
ИППИ РАН, МФТИ (ГУ)
vasilyuk@phystech.edu

Королев С.О.
ИППИ РАН, Сколтех
skorolev05@gmail.com

Ткачев А.И.
ИППИ РАН
annatkachev42@gmail.com

Беляев М. Г.
Сколтех, ИППИ РАН
belyaevmichel@gmail.com

Аннотация

В работе было произведено исследование алгоритмов обработки хромато-масс-спектрограмм для решения задач классификации. Был изучен вопрос возможности разделения отдельных липидов и способы решения задачи выравнивания хромато-масс-спектрограмм инструментами машинного обучения. Был создан алгоритм генерации признаков на основе частично обработанных данных для последующей классификации.

Для оценки качества работы алгоритма, были построены классификаторы на базе спектров 485 образцов липидного состава тканей головного мозга, позволяющие диагностировать такие заболевания, как аутизм и шизофрения. Качество классификации на основе предложенных признаков превосходит качество моделей, построенных с помощью стандартных алгоритмов выравнивания спектров.

Ключевые слова: липидомика, хромато-масс-спектрометрия, выравнивание пиков, кластеризация

1. Введение

Липидомика – активно развивающаяся область биологии, изучающая поведение липидов – нерастворимых органических молекул. Главной задачей является качественное разделение образца на составляющие липиды. Знание липидного состава предоставляет широкие возможности:

• Коммерческое использование. Оценка качества продовольственных продуктов, выявление подделок. Аутентификация[1].

• Фундаментальное исследование патологий. Изучение липидного состава тканей в зависимости от возраста. Диагностирование расстройств.

• Разработка пищевых добавок для коррекции метаболизма человека

Существует целый ряд методик, позволяющих определить качественно и количественно содержание липидов в образце. К ним относятся: ядерный магнитный резонанс (NMR), масс-спектрометрия (MS), колебательная спектроскопия (vibrational spectroscopy).

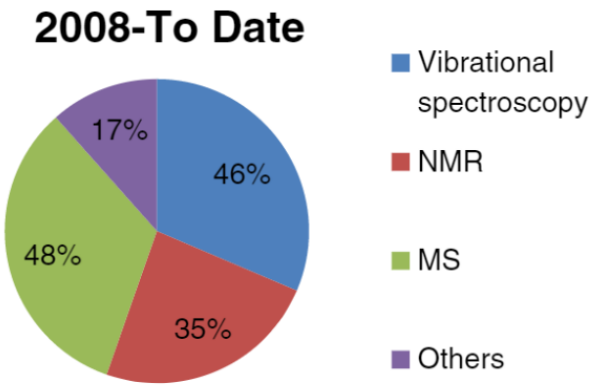

Рис. 1: Опубликованные работы в области изучения метаболитов[1].

Неотъемлемой частью исследования липидного состава при помощи масс-спектрометра является хроматограф. Он позволяет, основываясь на различных свойствах молекул, таких как гидрофобность, гидрофильность, способность растворяться или адсорбироваться, разделить молекулы во времени так, что в каждый момент на детектор попадает лишь небольшая часть из общей массы липидов.

Используемые хроматографы в зависимости от подвижной фазы разделяют на жидкостные и газовые. Газовые хроматографы (GC) способны работать только с летучими и термически устойчивыми веществами. Благодаря большой постоянности и силе разделения газовые хроматографы являются основным инструментом в задачах аутентификации продуктов питания. Жидкостные хроматографы (LC) требуют больших затрат времени и применяются главным образом в медицинских задачах[1]. Тандемная масс-спектрометрия (MS-MS) представляет собой 2 масс-спектрометра, соединенные так, что первый отсеивает несколько молекул, пришедших из хроматографа с наибольшей интенсивностью, а второй расщепляет эти молекулы, позволяя получить на выходе их спектры. Однако MS-MS системы способны обрабатывать только 10-20% информации, приходящей в первый масс-спектрометр[2].

Хромато-масс спектрограмма представляет собой трехмерный график зависимости интенсивности от времени запаздывания и отношения массы к заряду. Время запаздывания – время, которое уходит у молекулы на прохождение хроматографа. Следует отметить, что оно сильно зависит как от хроматографа, так и от материала, используемого в нем. Тем не менее, знание времени запаздывания может существенно упростить идентификацию липидов[3].

Главной задачей предварительной обработки данных с хромато-масс-спектрометра является представление их в виде, позволяющем легко получать доступ к таким свойствам отдельных ионов, как их время задержки, отношение заряда к массе и интенсивность.

Типичная последовательность шагов предварительной обработки состоит из фильтрации, обнаружении признаков, выравнивании по времени задержки и нормализации. Фильтрация позволяет избавиться от шума и влияния ограничений динамического диапазона на интенсивности. Обнаружение признаков позволяет выделить отдельные ионы из необработанного сигнала. Выравнивание объединяет измерения из нескольких образцов, а нормализация убирает нежелательные изменения интенсивности[4].

В силу своих химических свойств, различные молекулы ионизируются при разной поляризации: фосфолипиды и сфинголипиды ионизируются при отрицательном напряжении, ацилглицеролы - при положительном. Это делает необходимым проводить для каждого образца две прогонки в масс-спектрометре[5].

2. Постановка задачи и исходные данные

Задача. Исследовать возможность разделения липидов при обработке данных. Придумать алгоритм для создания признаков на основе хромато-масс-спектрограмм. Проверить возможность выравнивания пиков средствами анализа данных. Создать классификаторы на основе этих признаков и оценить их эффективность.

Определить, какие зависимости можно получить из имеющихся данных.

Данные. 485 LC масс-спектров образцов головного мозга в положительной и отрицательной ионизациях. Отношение массы к заряду фиксировано в пределах от 150 до 1500 m/z. Время измерения находится в пределах 20 минут, с точностью до одной секунды. В среднем, каждую секунду наблюдается 250 пиков. Минимальное расстояние между пиками $15 \cdot 10^{-5} m/z$. Среднее расстояние между пиками $5 \cdot 10^{-3} m/z$.

Молекулы не распределяются во времени равномерно.

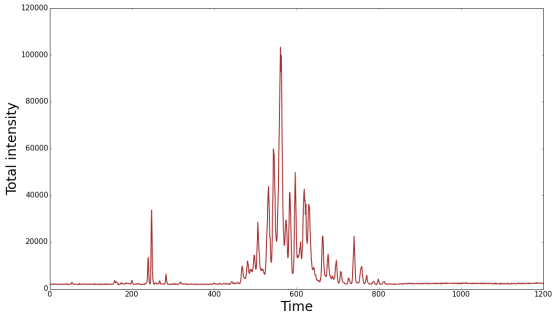

Рис. 2: Зависимость полной интенсивности ионного тока от времени.

3. Особенности данных

1. Принцип работы масс-спектрографа основан на движении заряженных частиц в магнитном поле, что позволяет определить массу не в чистом виде, а только как отношение массы к заряду. Таким образом, всегда существует некоторое количество n-раз кратно заряженных молекул.

2. При ионизации молекулы могут распадаться, поэтому одна молекула способна создавать целые спектры m/z.

3. В природе около 1% углерода встречается в виде его изотопа ^{13}C. Таким образом, молекулы с

числом атомов углерода > 50 имеют побочный пик с интенсивностью не менее 50% от интенсивности главного пика.

4. Масс-спектрометр имеет ограниченный динамический диапазон, что сказывается при большом потоке молекул.

5. В образцах всегда присутствует некоторая непостоянность: как в серии проб из одного источника, так и в анализах, произведенных в разных лабораториях. Значения интенсивностей и времена задержки меняются в силу физических свойств и невозможности точно повторить размер образца.

6. Разные хроматографы или хроматографы, отличающиеся подвижным носителем, могут давать отличающиеся разделения молекул, которые невозможно совместить аффинным преобразованием.

7. При больших энергиях ионизации молекулы распадаются. Для разных энергий ионизации спектры могут кардинально отличаться.

8. Время запаздывания можно считать мало меняющимся. При измерении времени запаздывания на протяжении 18 месяцев для разных типов клеток или тканей, его различия не превышали 1,25% [5].

4. Исследовательский процесс

В первую очередь была произведена попытка разделить липиды, чьи пики появлялись одновременно. Согласно гипотезе, если бы разные липиды имели различные фронты распространения в жидкой среде, то и производные их интенсивностей по времени отличались. Тем временем различные пики от одной молекулы должны иметь пропорциональные производные, так как процесс их образования является вероятностным, и отношения их интенсивностей определяют эти вероятности.

Для проверки гипотезы была построена двухмерная сетка с масштабом в 20 точек на единицу m/z, масштаб во времени сохранен. После этого каждый исходный пик отображался в некоторую его окрестность и итоговая интенсивность на сетке рассчитывалась по формуле [6]:

$$int_d(x,t) = \sum_{\substack{i=1 \\ t_i=t}}^{N} I_i \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2\sigma^2}} \qquad (1)$$

где I_i - исходная интенсивность i-го пика, x_i - его m/z, t_i - его время запаздывания, int_d - интенсивность на сетке, x - m/z на сетке, t - время запаздывания на сетке, σ - параметр, определяющий ширину отображенного пика, N - число пиков в исходных данных.

Путем наблюдения за интенсивностью отдельных масс было установлено, что характерное время образования пика – 3 единицы времени. Таким образом, к сетке были последовательно применены операции фильтрации шума, логарифмирования и дифференцирования по трем точкам.

Как оказалось, разницы в значениях производной недостаточно для надежной классификации. Предположительно, это свидетельствует о том, что форма фронта является общей для разных молекул.

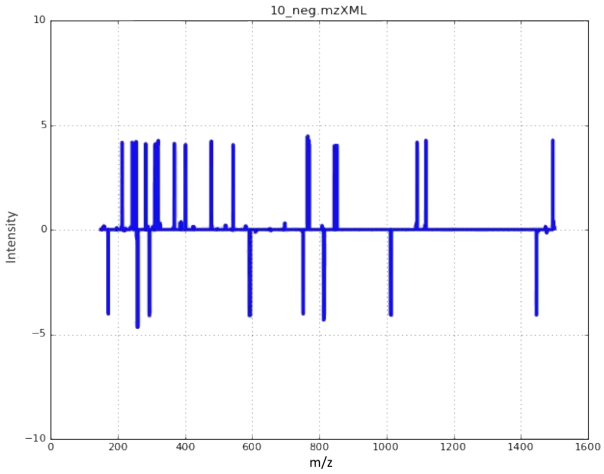

Рис. 3: Значения производной по времени от логарифма интенсивности в один из моментов времени.

Дальнейшей целью исследования было решение задачи классификации. Среди известных метаданных были возраст, пол, заболевание (человек мог быть здоровым, либо аутистом, либо шизофреником, либо иметь синдром Дауна) и раса человека, а также страна происхождения данных. Для этого был разработан алгоритм генерации признаков на основе хромато-масс-спектрограмм. Были построены классификаторы "каждый-против-каждого" и "один-против-всех" [7] для таких характеристик, как заболевание, раса, пол и исходная лаборатория.

5. Предобработка

При распаде или ионизации молекулы, её состав может измениться только на целое число протонов или нейтронов, массы которых близки к 1 а. е. м. Таким образом, возможно приближенно судить о молекуле по её массе, округленной до целого числа. Вкупе с отсутствием информации о точности исходных данных, было принято решение о переходе к целочисленным значениям m/z. Каждый пик распределялся среди двух ближайших целых m/z с весами, пропорциональными расстоянию до них.

При этом было невозможно суммировать интенсивности разных пиков, так как не было информации об исходной форме спектра, и возможна ситуация, когда суммирование даст заведомо неверный результат. Как видно из Рис. 4, при несглаженных исходных данных, алгоритмы могут находить несколько пиков на месте одного реального. Вкупе с тем, что среднее расстояние между имеющимися пиками составляет $5 \cdot 10^{-2} m/z$, можно считать, что эта ситуация имеет место быть.

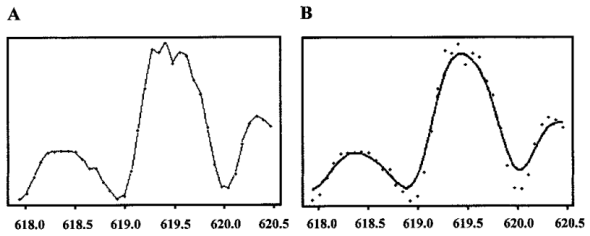

Рис. 4: Информация, полученная с масс-спектрометра в диапазоне 618-620.5 m/z до и после сглаживания[8].

Если просуммировать все пики в окрестности некоторой массы, то мы, возможно, получим результат в разы отличающийся от реального. Поэтому, для подсчета интенсивности в точке, был подсчитан максимум среди интенсивностей в некоторой окрестности.

Для осуществления выравнивания пиков по времени, было принято решение найти все пики в имеющихся хромато-масс-спектрограммах и кластеризовать их. Это позволило соотнести каждому регулярно встречающемуся пику целую группу пиков, соответствующих одному иону.

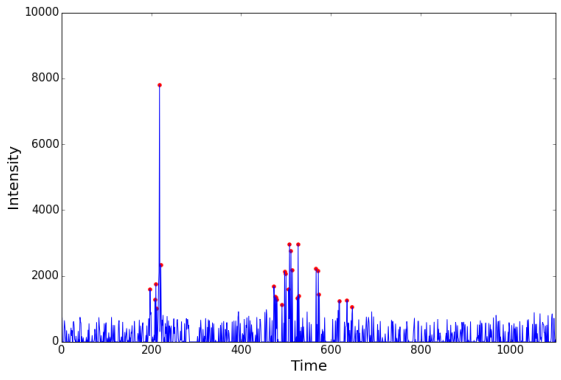

Рис. 5: Зависимость интенсивности от времени при $m/z = 260$.

Для этого находились пики интенсивности каждой массы во времени, и среди них отбирались только те, которые превосходили средний уровень интенсивности более, чем в 5 раз (величина получена эвристически). Это позволило снизить объем данных, получаемых от одного человека.

После получения положений пиков со всех образцов для одной поляризации, мы подсчитали, сколько пиков приходилось на каждую точку плоскости <m/z, время запаздывания>, и оставили только те точки, в которых пики встречались чаще, чем в среднем, в e раз.

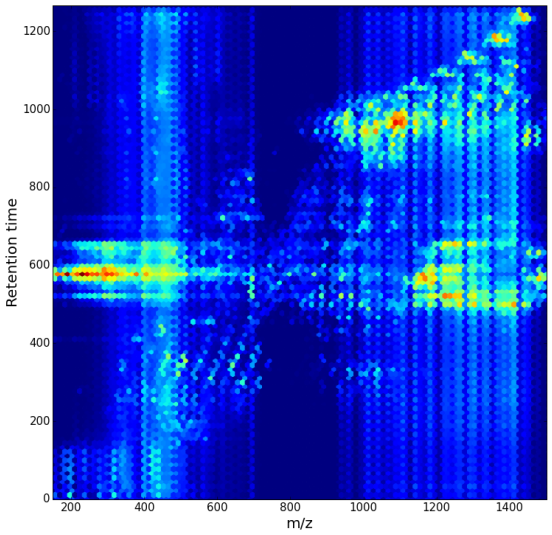

Рис. 6: Частота появления пиков среди всех хромато-масс-спектрограмм.

Чтобы создать отдельные признаки для дальнейшей классификации, был использован алгоритм DBSCAN с евклидовой метрикой, но уменьшенным масштабом по времени.

В итоге было получено приблизительно 1500 кластеров для каждой поляризации, среди которых более 500 имеют более 5 точек.

Listing 1 Алгоритм создания шаблона из кластера точек

```
template = {};
for point in cluster:
   mass, time = point
   if template.has_key(mass):
      min_, max_ = template[mass]
      min_ = min(min_, time)
      max_ = max(max_, time)
      template[mass] = (min_, max_)
   else:
      template[mass]=(time,time)
```

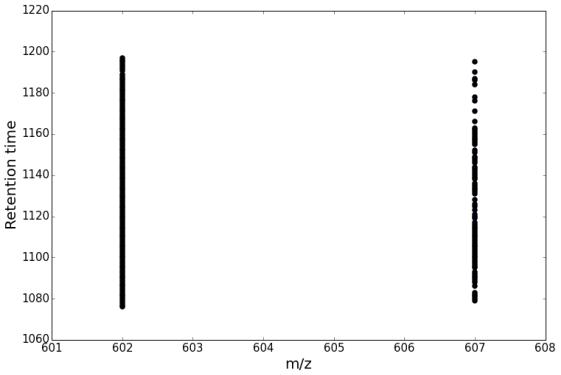

Рис. 7: Пример кластера.

Из кластера создавался шаблон, который определял признак для классификации. Для этого, для каждой присутствующей массы выбирались минимум и максимум по времени, они дополнительно увеличивались на 3 единицы времени чтобы придать бóльшую стабильность методу, массив из них сохранялся как шаблон. Также это позволило восстановить пропущенные точки в середине кластера.

6. Результаты

Оценка предложенного способа генерации пиков проводилась при помощи нескольких широко распространенных алгоритмов классификации:

- Градиентный бустинг на решающих деревьях, реализация XGBoost [9];
- логистическая регрессия (в таблицах далее обозначается как LogReg), реализация scikit-learn [10];

При помощи логистической регрессии результаты получались на основе описанных выше признаков. В это же время, на вход XGBoost подавалось либо отношение признаков к сумме всех интенсивностей образца, либо отношение признаков к признаку, выбранному для данной задачи классификации.

Для оценки качества классификации была организована кросс-проверка с помощью 5 стратифицированных разбиений. Метрикой качества классификации служила площадь под ROC-кривой (ROC_AUC [11]). В качестве базового метода генерации признаков, с которым проводилось сравнение предложенного подхода, использовались стандартные алгоритмы выравнивания из программной библиотеки MZmine 2 [12], предназначенной для обработки хромато-масс-спектров.

6.1. Сравнение со стандартными подходами к выравниванию пиков

Для всех исследованных постановок задач классификации на данной выборке спектров, полученные предложенным алгоритмом признаки позволяют добиться более высокого качества классификации по сравнению с классификацией на основе стандартных подходов с выравниванием пиков.

Здоровый-Все	Шизофрения-Здоровый	Аутизм-Шизофрения	Здоровый-Аутизм
0.80 ± 0.05	0.86 ± 0.02	0.94 ± 0.03	0.81 ± 0.04
0.75 ± 0.08	0.88 ± 0.05	0.92 ± 0.06	0.78 ± 0.02

Таблица 1: Сравнительная таблица точности классификации. Верхняя строчка - площадь под ROC-кривой классификаторов, которые использовали предложенный подход для генерации признаков, нижняя - аналогичная метрика для классификаторов, которые были построены на основе признаков, созданных существующими алгоритмами выравнивания пиков.

6.2. Качество классификации в различных задачах

Помимо сравнения с общепринятым подходом к генерации признаков в основной задаче классификации, мы также использовали метаданные для формулировки нескольких дополнительных задач классификации и оценили точность решения этих задач предложенным методом.

Наилучшее качество классификации среди всех поставленных задач было получено при отделении азиатов или китайской лаборатории среди всех записей. Так как в данных эти свойства представляют одни и те же записи, мы не могли определить, какое именно свойство имело ярко выраженные особенности.

Классифицировать мужчин и женщин на основе имеющихся признаков не получилось. Ни один алгоритм не дал точность классификации ROC_AUC выше 0.65.

В то же время качества классификации по заболеванию достаточно для выявления зависимости. Лучше всего в данном случае удалось диагностировать шизофрению.

Помимо прочего, полученные признаки имеют четкую биологическую интерпретацию. Благодаря известной массе и времени задержки, каждый признак отсылает к определенному липиду. В дальнейшем можно применить MS-MS масс-спектрометрию, определив за какое непосредственно вещество отвечает каждый признак. Также можно направленно измерять его концентрацию в веществе. Из имею-

Задача	XGBoost	LogReg
Шизофрения-Все	0.88 ± 0.02	0.88 ± 0.02
Аутизм-Все	0.78 ± 0.04	0.75 ± 0.06
Здоровый-Все	0.80 ± 0.02	0.76 ± 0.02
Шизофрения-Аутизм	0.89 ± 0.02	0.94 ± 0.03
Шизофрения-Здоровый	0.86 ± 0.03	0.86 ± 0.02
Здоровый-Аутизм	0.79 ± 0.05	0.81 ± 0.04
Азиат-Все	0.98 ± 0.007	0.94 ± 0.016
Мужчина-Женщина	0.64 ± 0.02	0.65 ± 0.02

Таблица 2: ROC_AUC построенных классификаторов.

щихся данных определить реальное количество веществ возможным не представляется, так как неизвестна ионизирующая способность прибора для выбранных молекул.

7. Дальнейшее развитие

Среди созданных кластеров имеются около 10 для каждой поляризации с чрезмерно большим числом точек. При этом в списки самых значимых признаков они не входят. Модифицированный алгоритм выделение кластеров, разбивающий эти крупные кластеры на подмножества, мог бы позволить получить более высокое качество классификации. В данной работе использовались центроидные данные, то есть был проведен первый этап предварительной обработки сырых спектров - выделены отдельные пики. При работы с неоработанными данными можно более качественно перейти к сетке с целыми значениями m/z.

8. Выводы

В работе было произведено исследование возможности разделения липидов, основываясь на скорости роста их интенсивностей. Создан алгоритм генерации признаков, основанный на кластеризации пиков в данных хромато-масс-спектров. Эффективность предложенного алгоритма была продемонстрирована с помощью решения задачи классификации шизофрении и аутизма по результатам хромато-масс-спектрометрии тканей головного мозга.

Список литературы

[1] Elena Cubero-Leon, Rosa Peñalver, and Alain Maquet. Review on metabolomics for food authentication. *Food Research International*, 60:95–107, 2014.

[2] Rob Smith, Andrew D Mathis, Dan Ventura, and John T Prince. Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC bioinformatics*, 15(7):1, 2014.

[3] Fabian Aicheler, Jia Li, Miriam Hoene, Rainer Lehmann, Guowang Xu, and Oliver Kohlbacher. Retention time prediction improves identification in nontargeted lipidomics approaches. *Analytical chemistry*, 87(15):7698–7704, 2015.

[4] Mikko Katajamaa and Matej Orešič. Data processing for mass spectrometry-based metabolomics. *Journal of chromatography A*, 1158(1):318–328, 2007.

[5] Laxman Yetukuri, Mikko Katajamaa, Gema Medina-Gomez, Tuulikki Seppänen-Laakso, Antonio Vidal-Puig, and Matej Orešič. Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. *BMC Systems Biology*, 1(1):1, 2007.

[6] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[7] David MJ Tax and Robert PW Duin. Using two-class classifiers for multiclass classification. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 124–127. IEEE, 2002.

[8] Jeffrey S Forrester, Stephen B Milne, Pavlina T Ivanova, and H Alex Brown. Computational lipidomics: a multiplexed analysis of dynamic changes in membrane lipid composition during signal transduction. *Molecular pharmacology*, 65(4):813–821, 2004.

[9] *XGBoost*. https://github.com/dmlc/xgboost.

[10] *Scikit-learn*. http://scikit-learn.org.

[11] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[12] Tomáš Pluskal, Sandra Castillo, Alejandro Villar-Briones, and Matej Orešič. Mzmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*, 11(1):1, 2010.