УДК 004.8

И.И. Чернявский, Ф. Александров, С.И. Николенко

МЕТОД СЕГМЕНТАЦИИ РЕЗУЛЬТАТОВ MALDI-СПЕКТРОМЕТРИИ НА ОСНОВЕ ГРАФИЧЕСКИХ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ

Чернявский И.И., Александров Ф., Николенко С.И. **Метод сегментации результатов MALDI-спектрометрии на основе графических вероятностных моделей.**

Аннотация. В работе рассматривается задача сегментации масс-спектрометрических изображений, полученных методом MALDI. Предлагается подход, основанный на применении графических моделей (модели LDA и марковских сетей) для решения задачи. Рассматриваются несколько модификаций подхода и проводится сравнение с известными решениями; выделяются преимущества предлагаемого подхода.

Ключевые слова: MALDI, масс-спектрометрия, сегментация, графическая модель, LDA, марковская сеть, модель Изинга.

Chernyavsky I.I., Alexandrov T., Nikolenko S.I. **Segmentation of MALDI imaging results based on graphical models.**

Abstract. We consider the MALDI imaging segmentation problem and propose an approach based on graphical models (LDA model and Markov random fields). We consider several modifications to the proposed approach and compare it to the previously known methods; we distinguish several advantages of the proposed approach.

Keywords: MALDI, mass spectrometry, segmentation, graphical model, LDA, Markov random field, Ising model.

1. Введение.

1.1. Методы масс-спектрометрического изображения.

Масс-спектрометрия – один из главных современных методов биохимического анализа. К методам масс-спектрометрии относятся, в частности, методы масс-спектрометрического изображения (*imaging mass spectrometry*); они активно применяются при изучении образцов тканей, разработке лекарств, поиске биомаркеров и для ряда других задач [28]. Различают несколько разновидностей изображающей масс-спектрометрии, основанных на разных физических принципах. Наиболее распространённым вариантом является метод матрично-активированной лазерной десорбции/ионизации (*Matrix Assisted Laser Desorption / Ionization – MALDI*), получивший распространение в конце 1990-х годов и применяющийся для анализа высокомолекулярных соединений (пептиды, белки и др.). [28].

В рамках метода MALDI на образец ткани наносится матрица для улучшения процесса ионизации. Затем образец помещается в масс-спектрометр, где при помощи лазера происходит ионизация вещества. Прибор определяет отношения m/z (масса/заряд) полученных частиц и составляет *спектр* – вектор интенсивностей для различных значений m/z. Результатом работы MALDI-спектрометра является набор спектров для каждой точки поверхности образца. Таким образом, результат можно рассматривать как «куб данных», гипер-изображение, каждому пикселю которого сопоставлен вектор большой размерности.

1.2. Задача сегментации. Задача сегментации результатов MALDI-спектрометрии заключается в определении структурно-однородных областей образца и их визуализации в виде сегментов на изображении. Данная задача представляет большой практический интерес, так как полученная картина сегментации показывает внутреннее строение образца. Кроме непосредственной картины сегментации, большой интерес также представляют характерные массы, выраженные в каждом сегменте. Информация о массах позволяет получить представление о биохимических различиях между найденными областями.

Существует несколько подходов к решению задачи сегментации. В следующих разделах мы рассмотрим известные методы сегментации на основе алгоритма кластеризации k-средних и предлагаемый в данной статье подход, основанный на применении графических моделей. Алгоритмы, рассматриваемые в данной работе, запускались на данных, полученных путем анализа среза мозга крысы в MALDI-спектрометре. Разрешение изображения составляет 201×120 пикселей. Измерялись интенсивности для 3045 значений масс в диапазоне 2.5-10кДа.

1.3. Результаты и структура статьи. Ранее сегментация изображений MALDI-спектрометрии проводилась методами, в основе которых не лежала вероятностная модель [1]. Хотя такие методы могут позволить получить сегментацию высокого качества, у них достаточно низкий потенциал для дальнейших обобщений, которые крайне интересны для задач обработки данных масс-спектрометрии. В частности, жёсткая сегментация плохо приспособлена для дальнейшего анализа полученных сегментов; но на самом деле нас как раз интересуют не столько сами сегменты, сколько их состав, например характерные для каждого сегмента

массы.

В этой работе мы предлагаем подход, ранее не применявшийся к сегментации MALDI-данных: вероятностную кластеризацию с последующим сглаживанием. Предобработка при помощи кластеризации нужна здесь для того, чтобы разделить крайне многомерные данные (спектр длины около 3000) в каждой точке изображения на похожие подмножества, тем самым, в частности, сократив размерность. Затем мы применяем метод сглаживания, основанный на ненаправленных графических моделях (марковских полях).

Работа организована следующим образом: в разделе 2 мы рассматриваем методы кластеризации, которые мы применяем в этой работе для анализа MALDI-данных: алгоритм k-средних и модель LDA; в разделе 3 мы вводим графические вероятностные модели и показываем, как ненаправленные графические модели (марковские случайные поля) применяются для сглаживания; раздел 4 посвящён общему описанию нашей модели обработки данных масс-спектрометрии; в нём приводятся и практические результаты на описанных выше данных; раздел 5 завершает работу.

2. Кластеризация.

2.1. Алгоритм кластеризации k-средних. Алгоритм k-средних является широко используемым итеративным алгоритмом кластеризации [14,16]. На вход алгоритму подаётся число кластеров k и точки из входного набора данных; на этом наборе должна быть задана метрика для определения расстояний (точки набора должны лежать в евклидовом пространстве, обычно $\mathbb{R}^n$). Алгоритм начинает работу со случайного выбора k векторов в качестве центров кластеров. Затем алгоритм повторяет следующие шаги:

- на основе текущих центров кластеров, определить для каждой точки ближайший к ней кластер;
- пересчитать центры кластеров, взяв за новый центр - среднее значение точек в кластере.

Алгоритм прекращает работу либо когда число итераций превысит заданное, либо пока кластеры не перестанут меняться. Алгоритм k-средних фактически является вариантом вероятностного EM-алгоритма для случая «жёсткой кластеризации», когда вместо численных значений правдоподобий принадлежности точек кла-

стерам алгоритм поддерживает и использует только индекс наиболее правдоподобного кластера.

2.2. Сегментация на основе алгоритма k-средних. Один из известных подходов к сегментации основан на прямом применении алгоритма k-средних к спектрам. На рис. 1 изображен результат применения алгоритма k-средних к спектрам, точки одного цвета соответствуют одному кластеру. Из рисунка видно, что полученная сегментация крайне зашумлена. Причиной шума являются неточные измерения масс-спектрометра и, как следствие, ошибки в данных.

Для того чтобы решить проблему ошибок в данных, в [1] предложен метод сегментации, основанный на пространственном сглаживании спектров, в котором спектр точки сглаживается с использованием спектров соседних точек, и применении к результату сглаживания алгоритма k-средних. Пример полученной сегментации изображен на рис. 2 [1]. Преимуществом данного подхода является использование пространственной структуры на спектрах и значительное повышение качества полученной сегментации. Однако данный подход не предоставляет дополнительной информации для последующего анализа; в частности, он не позволяет непосредственно получать характерные массы для каждого сегмента.

2.3. Модель LDA. Подход к сглаживанию изображений, получающихся при MALDI-спектрометрии, предлагаемый в данной работе, главным образом основан на применении модели LDA (Latent Dirichlet Allocation), изображённой на рис. 3 [7,17]. Изначально модель LDA формулировалась в терминах задачи обработки текстов (для последующего использования в информационном поиске, категоризации и др.). В рамках данной задачи необходимо проанализировать набор текстовых документов и вывести генеративное распределение, выделив основные темы, встречающиеся в документах.

LDA является *порождающей* моделью и может быть использована для определения тем каждого документа. При этом предполагается, что документ есть смесь нескольких тем, каждая из которых задается распределением на словах. Более формально, LDA задает следующий процесс порождения документа.

1. Выбрать число слов в документе: $N \sim p(N \mid \xi)$.
2. Выбрать распределение тем для документа: $\theta \sim Dir(\alpha)$.

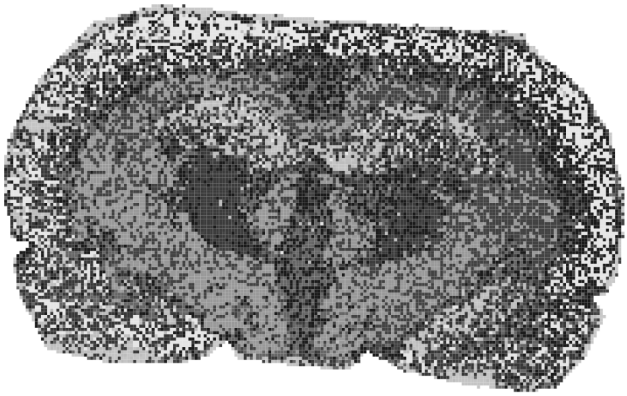

Рис. 1: Прямое применение алгоритма k-средних к спектрам

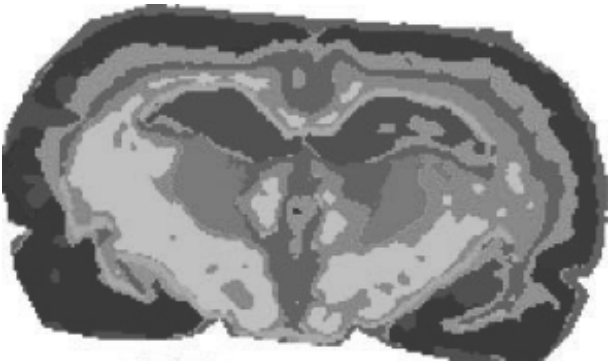

Рис. 2: Сегментация, полученная с помощью пространственного сглаживания и метода k-средних [1]

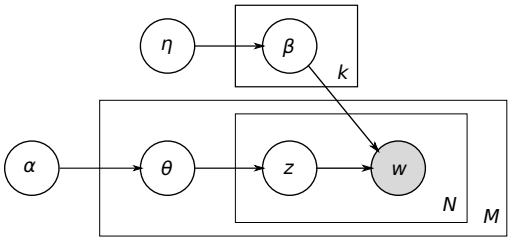

Рис. 3: Модель LDA

3. Для каждого слова w_n, $n = 1..N$:

   (a) выбрать тему из распределения документа: $z_n \sim \mathrm{Mult}(\theta)$;

   (b) выбрать слово из распределения темы z_n на словах: $w_n \sim p(w_n \mid z_n, \beta)$.

Заметим, что данная модель не учитывает порядок слов в документе, т.е. документ – это *bag of words*.

Зная, из каких слов состоят документы, можно получить максимальные апостериорные значения θ_d для каждого документа и $p(w \mid z, \beta)$ для каждой темы. В качестве алгоритма вывода можно использовать сэмплирование по Гиббсу, вариационные методы и другие алгоритмы.

Модель LDA хорошо зарекомендовала себя в области обработки текстов, но применения этой модели не ограничиваются только задачами информационного поиска. В частности, в этой работе мы предлагаем применить LDA для анализа спектров. Для этого необходимо ввести понятие «документа» и «слова» в документе. Предлагается в качестве документов рассмотреть отдельные пиксели изображения, полученного методом MALDI, а в качестве слов – значения масс, наиболее выраженных в спектре точки-пикселя (пики спектра).

Результатом применения LDA к таким «документам» являются распределения на темы в каждом пикселе изображения и распределение на массы для каждой темы. Число тем является параметром, определяемым заранее. На рис. 4 изображен результат прямого применения LDA с числом тем 10. Каждому цвету на рисунке

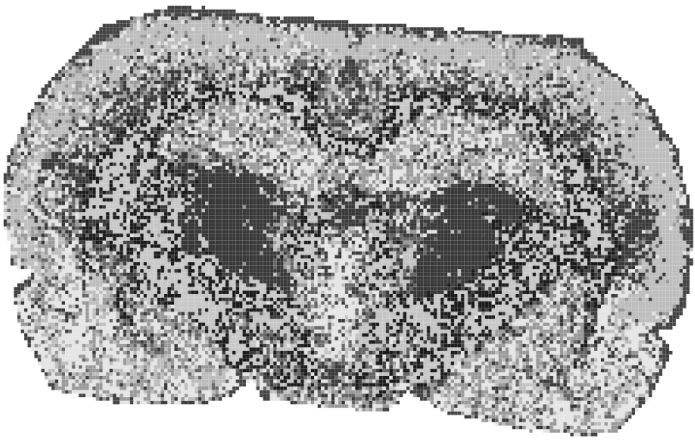

Рис. 4: Результат применения LDA

соответствует своя тема, для изображения в каждой точке выбиралась тема, имеющая наибольшую апостериорную вероятность.

3. Ненаправленные графические модели и вывод на них.

3.1. Марковские сети. В рамках описываемого в данной статье подхода используются ненаправленные графические модели, известные как *марковские сети* (Markov networks, Markov random fields) [4, 15, 19, 23, 34]. Марковские сети являются разновидностью графических моделей [4, 19, 34], которые представляют собой мощный инструмент машинного обучения и используются для эффективной работы с совместными распределениями большого числа случайных величин. Многие задачи можно свести к задачам поиска наиболее вероятных значений случайных переменных, поэтому графические модели нашли широкое применение; обычно графические модели делятся на марковские сети (ненаправленные) и байесовские сети доверия (направленные) [20, 21, 29, 34]; другой тип графических вероятностных моделей представляют собой алгебраические байесовские сети [30–37].

В общем случае совместное распределение n случайных вели-

чин имеет экспоненциальное от n число свободных параметров, поэтому работа с таким распределением напрямую невозможна. Однако распределения, встречаемые на практике, как правило, имеют свойства, позволяющие эффективно работать со случайными величинами. Данные свойства заключаются в имеющихся в распределениях условных независимостях. Графическая модель позволяет эффективно работать с распределением, выражая имеющиеся в распределении условные независимости в виде графа.

Марковская сеть является разновидностью графической модели, использующей ненаправленный граф для представления распределения. Введём основные определения теории марковских сетей.

Определение 1. Пусть $D = \{X_{i_1}, X_{i_2}, ..., X_{i_n}\}$ – множество случайных переменных; тогда функция $\phi(D)$ является *фактором*, если

$$\phi(D) : Val(D) \to \mathbb{R}_+,$$

где $Val(D)$ – декартово произведение множеств значений случайных переменных из D.

Определение 2. Рассмотрим множество случайных переменных $X = \{X_1, X_2, ..., X_n\}$; тогда совместное распределение $P(X)$ является *распределением Гиббса*, параметризованным множеством факторов $\Phi = \{\phi_1(D_1), \phi_2(D_2), \phi_m(D_m)\}$, если оно имеет следующий вид:

$$P(X_1, X_2, ..., X_n) = \frac{1}{Z}\phi_1(D_1)\phi_2(D_2)...\phi_m(D_m),$$
$$Z = \sum_{X_1, X_2, ..., X_n} \phi_1(D_1)\phi_2(D_2)...\phi_m(D_m).$$

Определение 3. Назовем формально *марковской сетью* ненаправленный граф, вершинам которого соответствуют случайные переменные $\{X_1, X_2, ..., X_n\}$.

Определение 4. Распределение Гиббса, параметризованное множеством факторов $\Phi = \{\phi_1(D_1), \phi_2(D_2), \phi_m(D_m)\}$, *факторизуется* по марковской сети H, если для всех $k \in \{1, ..., m\}$ множество вершин, соответствующих переменным D_k, является полным подграфом H.

Марковская сеть задаёт ряд утверждений об условных независимостях в распределении.

Определение 5. Множество переменных Z *отделяет* множество переменных X от множества переменных Y в марковской сети H, если все пути из X в Y содержат вершины из Z. Определим множество условных независимостей для марковской сети H следующим образом:

$$I(H) = \{(X \perp Y|Z) : Z \text{ отделяет } X \text{ от } Y\}.$$

Запись $(X \perp Y|Z)$ означает: «множество переменных X и множество переменных Y условно независимы при условии Z».

В следующих классических теоремах устанавливается важная связь между факторизацией по марковской сети и условными независимостями в распределении.

Теорема. *Если распределение $P(X)$ факторизуется по марковской сети H, то для него верны все свойства $I(H)$ об условной независимости.*

Теорема (Hammersley-Clifford). *Если для положительного (т.е. вероятность любого события строго больше 0) распределения $P(X)$ верны все утверждения $I(H)$ марковской сети H, то $P(X)$ факторизуется по марковской сети H.*

Таким образом, марковская сеть задает множество условных независимостей и класс распределений, удовлетворяющих данным независимостям. Данный класс – это в точности класс распределений, представимых в виде произведения факторов (деленного на нормализационную константу Z) над полными подграфами марковской сети.

3.2. Вывод в марковских сетях. Задача вывода на графической модели обычно ставится следующим образом: для данного совместного распределения $P(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m)$ найти апостериорное распределение $P(X_1, X_2, ..., X_n|Y_1, Y_2, ..., Y_m)$ для известных значений $Y_1, ..., Y_m$. Как правило, переменные $Y_1, ..., Y_m$ наблюдаются непосредственно и связаны с данными, переменные $X_1, ..., X_n$ – скрытые переменные, соответствующие неизвестным величинам, значения которых мы хотим оценить. Например, в задаче классификации переменные Y_i соответствуют наблюдаемым признакам объектов, X_j определяют принадлежность к тому или иному классу, а само распределение $P(X, Y)$ выражает зависимость между признаками объектов и соответствующими классами. Часто достаточно определить только наиболее вероятные значения $(x_1, x_2, ..., x_n)$ переменных $\{X_1, X_2, ..., X_n\}$ для апостериор-

ного распределения $P(X_1, X_2, ..., X_n|Y_1, Y_2, ..., Y_m)$. Такое множество значений называется максимальным апостериорным решением (*maximum a posteriori*, MAP-решение).

В общем случае задача вывода на графических моделях NP-полна. Однако факторизация совместного распределения многих переменных в виде произведения факторов, каждый из которых зависит только от небольшого числа переменных, позволяет найти эффективные алгоритмы для этой задачи. Существуют различные алгоритмы вывода на графических моделях. Рассмотрим простой, но мощный и активно применяющийся метод приближённого вывода – *сэмплирование по Гиббсу* [4, 8, 9, 24].

Предполагается, что значения переменных $\{Y_1, ..., Y_m\}$ известны и равны соответственно $(y_1, ..., y_m)$. Алгоритм сэмплирования по Гиббсу делает серию итераций, в конце каждой из которых выводятся значения переменных $\{X_1, ..., X_n\}$. Сначала вектор значений переменных $\{X_1, ..., X_n\}$ инициализируется случайными значениями $(x_1^0, x_2^0, ..., x_n^0)$. На каждой итерации алгоритм обновляет значения переменных по следующим формулам:

$$\begin{aligned}
x_1^{s+1} &\sim P(x_1|x_2^s, x_3^s, ..., x_n^s, y_{1..m}),\\
x_2^{s+1} &\sim P(x_2|x_1^{s+1}, x_3^s, ..., x_n^s, y_{1..m}),\\
&...\\
x_n^{s+1} &\sim P(x_N|x_1^{s+1}, x_2^{s+1}, ..., x_{n-1}^{s+1}, y_{1..m}).
\end{aligned}$$

Главным свойством данного случайного процесса является то, что после достаточно большого числа итераций распределение на получаемых в конце каждой итерации значениях $(x_1^s, x_2^s, ..., x_n^s)$ стремится к апостериорному $P(X_{\{1..n\}}|Y_{\{1..m\}})$.

Таким образом, алгоритм фактически позволяет сэмплировать из апостериорного распределения. Обычно первые N возвращаемых значений переменных $\{X_1, ..., X_n\}$ пропускают, а все последующие используют в качестве сэмплов из апостериорного распределения. Имея достаточно большое число сэмплов, можно надёжно аппроксимировать апостериорное распределение.

Представление совместного распределения в виде произведения факторов позволяет эффективно производить обновление значений переменных. Так, например, в марковских сетях, где все факторы являются функциями одной или двух переменных (*pairwise*

Markov networks – этот тип марковской сети мы будем использовать в нашем методе сглаживания), $P(X_i|X_{-i})$ можно легко вычислить:

$$P(x_i = l|x_{-i}) = \frac{P(x_i = l, x_{-i})}{\sum_k P(x_i = k, x_{-i})} =$$
$$= \frac{(1/Z)(\prod_{j\in N_i} \phi_{ij}(x_i = l, x_j))(\prod_{<jt>:j,t\notin N_i} \phi_{jt}(x_j, x_t))}{(1/Z)\sum_k(\prod_{j\in N_i} \phi_{ij}(x_i = k, x_j))(\prod_{<jt>:j,t\notin N_i} \phi_{jt}(x_j, x_t))} =$$
$$= \frac{\prod_{j\in N_i} \phi_{ij}(x_i = l, x_j)}{\sum_k \prod_{j\in N_i} \phi_{ij}(x_i = k, x_j)},$$

где N_i – множество номеров вершин, соседних с вершиной i в графе, а запись X_{-i} означает «все переменные, кроме i-й». Из формул видно, что для обновления значения переменной X_i, достаточно рассмотреть только значения в соседних вершинах графа.

3.3. Модель Изинга. Модель Изинга сначала возникла в статистической физике для изучения ферромагнитных свойств вещества. С точки зрения графических моделей модель Изинга является частным случаем марковской сети. Граф модели изображён на рис. 5. Граф модели является решёткой, вершины которой соответствуют случайным величинам, принимающим значения $\{+1, -1\}$. Распределение на случайных переменных в модели Изинга имеет следующий вид:

$$P(x_1, x_2, ..., x_n) = \frac{1}{Z} \exp\left(\sum_{i<j} w_{i,j} x_i x_j + \sum_i u_i x_i\right),$$

где $w_{i,j}$ и u_i – веса на рёбрах и в вершинах. С физической точки зрения каждой вершине соответствует атом в кристаллической решетке вещества, значениям $\{+1, -1\}$ – направление спина атома. Величина

$$-\sum_{i<j} w_{i,j} x_i x_j - \sum_i u_i x_i$$

соответствует энергии системы. Таким образом, с физической точки зрения состояние с наибольшей вероятностью имеет наименьшую энергию.

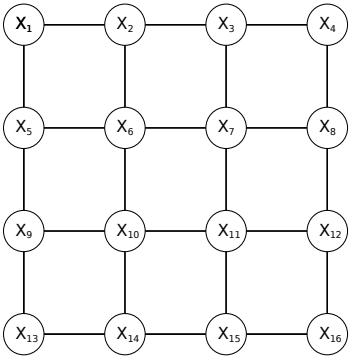

Рис. 5: Граф модели Изинга

Факторы вида $\exp(w_{i,j}x_ix_j)$ соответствуют каждому ребру и, при условии $w_{i,j} > 0$, принимают наибольшие значения, когда $x_i = x_j$. Таким образом, данные факторы способствуют равенству значений в соседних вершинах. Факторы вида $\exp(u_ix_i)$ соответствуют каждой вершине и, в зависимости от знака u_i, определяют более свойственное данной вершине значение. Примеры сэмплов из распределения модели Изинга приведены на рис. 6.

3.4. Сглаживание изображений на основе марковской сети. В задаче сглаживания изображений требуется удалить шум с изображения. Марковские сети – один из классических инструментов для решения данной задачи [18,22,23]; на самом деле, удаление шума с изображений было одним из первых прямых применений ненаправленных графических моделей [2,3,9].

Подобно модели Изинга, граф марковской сети имеет вид решётки. При этом бинарные факторы на рёбрах параметризованы параметром w, определяющим степень сглаживания, а унарные факторы в вершинах способствуют сохранению деталей изображения – наибольшее значение фактора в вершине достигается на значении, равном исходному значению в изображении с шумом. С помощью сэмплирования по Гиббсу можно найти наиболее вероятные значения пикселов и восстановить изображение.

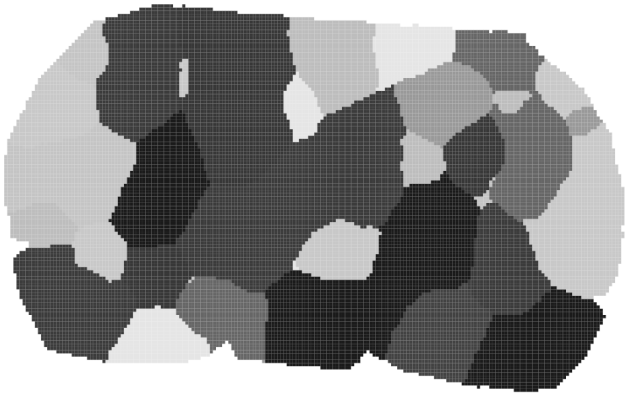

Рис. 6: Пример сэмплирования из модели Изинга

4. Обработка результатов MALDI-спектрометрии.

4.1. Сглаживание результатов кластеризации по методу k-средних. Как уже говорилось выше, марковские сети можно применить для сглаживания изображений любой природы, в частности, для сглаживания результатов MALDI-спектрометрии, предварительно сегментированных каким-либо другим способом. Рассмотрим для сравнения способ сглаживания с помощью марковских сетей, основанный на использовании алгоритма k-средних. Результатом работы алгоритма k-средних с параметром k являются k найденных центров кластеров. Рассмотрим марковскую сеть, аналогичную марковской сети в виде решётки из предыдущего пункта. Множество значений переменных здесь тоже составляет $\{1, 2, .., k\}$. Бинарные факторы $\phi_{i,j}(t_1, t_2)$, соответствующие рёбрам, параметризуются параметром w (вес) и имеют значение $\exp(w)$, если $t_1 = t_2$, и $\exp(-w)$, если $t_1 \neq t_2$. Унарные факторы обратно пропорциональны расстоянию от спектра до центра кластера:

$$u_i(t) = \frac{1}{||s_i - c_t||},$$

где s_i – спектр точки i, c_t – центр t-го кластера, найденный по методу k-средних. На рис. 7 приведены сглаженные изображения

для различных значений w.

4.2. Сглаживание результатов кластеризации при помощи модели LDA. Метод сегментации, предлагаемый в данной работе, основан на применении модели LDA и марковской сети для обработки и сглаживания результатов MALDI-спектрометрии.

Метод состоит из двух этапов. На первом этапе к спектрам применяется модель LDA, и, таким образом, результатом первого этапа являются распределения на темах в каждой точке и распределения на массах для каждой темы. На втором этапе происходит сглаживание картины сегментации с использованием марковской сети. Множество значений переменных, стоящих в отдельных пикселях, в данном случае $\{1, 2, .., k\}$, где k – число тем. Бинарные факторы $\phi_{i,j}(t_1, t_2)$, соответствующие рёбрам решетки, параметризуются параметром w (вес) и имеют значение $\exp(w)$, если $t_1 = t_2$, и $\exp(-w)$, если $t_1 \neq t_2$. Унарный фактор для точки i имеет следующий вид:

$$u_i(t) = \theta_i(t),$$

где θ_i – распределение на темы в данной точке, $\theta_i(t)$ – вероятность темы t. Таким образом, в данном методе для каждой точки выбирается её тема с учётом вероятностей тем в данной точке, полученных при помощи LDA, и значений тем в соседних точках.

Полученные сегментации для различных значений w изображены на рис. 8.

5. Заключение. В этой работе мы представили новый метод вероятностной обработки данных масс-спектрометрического изображения, полученных методом MALDI. Предлагаемый метод состоит из двух частей, каждая из которых представляет собой вывод на графической вероятностной модели: сначала кластеризация исходных спектров методом LDA, позволяющим выделить скрытые факторы («темы») отдельных кластеров, а затем сегментация полученного кластеризованного изображения при помощи ненаправленных графических моделей (случайных марковских полей). В дальнейшем мы планируем продолжать работу над улучшением результатов сегментации: улучшить модель LDA, обучая иерархические структуры и/или коррелированные темы [5, 6], расширить ненаправленную модель сглаживания скрытыми факторами, переходя таким образом от марковской сети к машине Больцмана (restricted Boltzmann machine) [10–13], применить алгоритмы сегментации, основанные на иерархических процессах Дирихле и про-

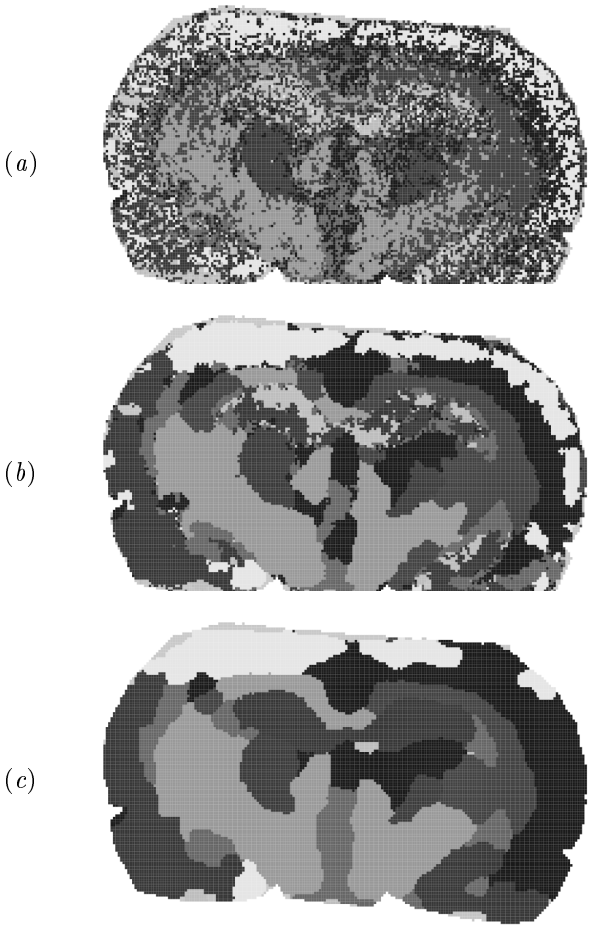

Рис. 7: Сглаживание с помощью метода k-средних и марковской сети: (a) $w = 0.3$; (b) $w = 0.6$; (c) $w = 0.9$

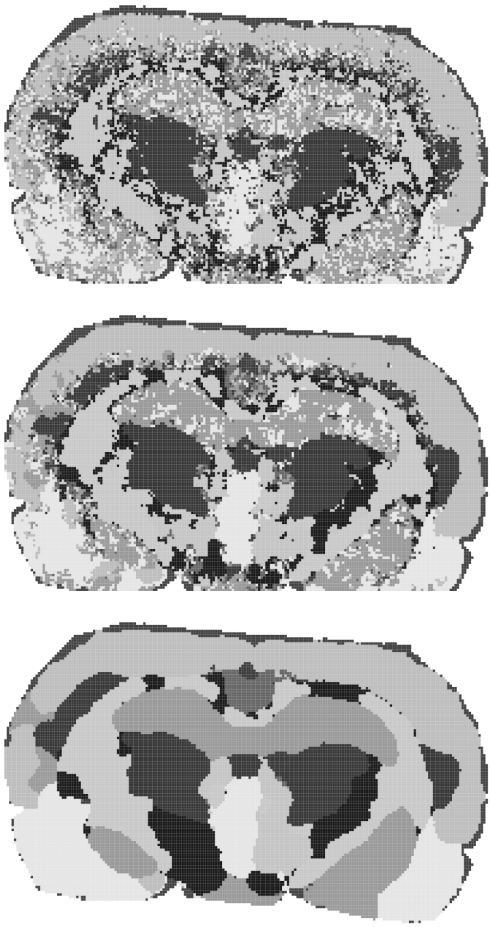

Рис. 8: Сегментация с помощью LDA и марковской сети (a) $w = 0.4$; (b) $w = 0.55$; (c) $w = 0.7$

цессах Питмана–Йора (hierarchical Dirichlet processes, hierarchical Pitman-Yor processes) [25–27]. Мы также планируем сделать следующий шаг в анализе «куба данных спектрометрии» и применить полученные модели для более глубокого анализа данных масс-спектрометрии: выделить массы, наиболее характерные для каждого сегмента, проанализировать масс-спектры сегментов, построить наиболее и наименее характерные срезы по массам.

Литература

1. *Alexandrov T., Kobarg J. H.* Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering // Bioinformatics. 2011. Vol. 27. P. i230–i238.

2. *Besag J.* On spatio-temporal models and Markov fields // Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes. Academia, 1974. P. 47–75.

3. *Besag J.* On the statistical analysis of dirty pictures // Journal of the Royal Statistical Society. 1986. Vol. B-48. P. 259–302.

4. *Bishop C. M.* Pattern Recognition and Machine Learning. Springer, 2006.

5. *Blei D. M., Jordan M. I., Griffiths T. L., Tennenbaum J. B.* Hierarchical Topic Models and the Nested Chinese Restaurant Process // Advances in Neural Information Processing Systems. 2004. Vol. 13.

6. *Blei D. M., Lafferty J. D.* Correlated topic models // Advances in Neural Information Processing Systems. 2006. Vol. 18.

7. *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. Vol. 3, N. 4–5. P. 993–1022.

8. *Casella G., George E. I.* Explaining the Gibbs sampler // The American Statistician. 1992. Vol. 46. P. 167–174.

9. *Geman S., Geman D.* Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984. Vol. 6. P. 721–741.

10. *Hinton G. E.* Training Products of Experts by Minimizing Contrastive Divergence // Neural Computation. 2002. Vol. 14. P. 1771–1800.

11. *Hinton G. E.* What kind of a graphical model is the brain? // Proceedings of the 19th International Joint Conference on Artificial Intelligence. 2005. P. 1765–1775.

12. *Hinton G. E.* A practical guide to training restricted Boltzmann machines. http://www.cs.toronto.edu/ hinton/absps/guideTR.pdf. 2012.

13. *Hinton G. E., Osindero S., Teh Y.-W.* A fast learning algorithm for deep belief nets // Neural Computation. 2006. Vol. 18. P. 1527–1554.

14. *Hudak P., Hughes J., Jones S. P., Wadler P.* Some Methods for classification and Analysis of Multivariate Observations // Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press, 1967. P. 281–297.

15. *Jordan M. I., Weiss Y.* Graphical models: probabilistic inference // Handbook of Neural Networks and Brain Theory / Ed. by M. Arbib. MIT Press, 2002.

16. *Kanungo T., Mount D. M., Netanyahu N., Piatko C., Silverman R., Wu A. Y.* An Efficient k–means Clustering Algorithm: Analysis and Implementation // IEEE Trans. Pattern Analysis and Machine Intelligence. 2002. Vol. 24. P. 881–892.

17. *Lacoste-Julien S., Sha F., Jordan M. I.* Disclda: Discriminative learning for dimensionality reduction and classification // Advances in Neural Information Processing Systems 20 / Ed. by J. Platt, D. Koller, Y. Singer, S. Roweis. Cambridge, MA: MIT Press, 2008.

18. *Li S. Z.* Markov Random Field Modeling in Image Analysis. Advances in Pattern Recognition. Springer, 2009.

19. *MacKay D. J.* Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003.

20. *Pearl J.* Probabilistic reasoning using graphs // Uncertainty in Knowledge-Based Systems / Ed. by B. Bouchon, R. R. Yager. Springer-Verlag, 1987. P. 201–202.

21. *Pearl J.* Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. NY etc.: Morgan Kaufmann, 1994.

22. *Perez P.* Markov random fields and images // CWI Quarterly. 1998. P. 413–437.

23. *Prince S.* Computer Vision: Models, Learning, and Inference. Cambridge University Press, 2012.

24. *Robert C. P., Casella G.* Monte Carlo Statistical Methods. New York: Springer-Verlag, 2004.

25. *Shyr A., Darrell T., Jordan M. I., Urtasun R.* Supervised hierarchical Pitman-Yor process for natural scene segmentation // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2011. P. 2281–2288.

26. *Sudderth E., Jordan M. I.* Shared segmentation of natural scenes using dependent Pitman-Yor processes // Advances in Neural Information Processing Systems 20 / Ed. by J. Platt, D. Koller, Y. Singer, S. Roweis. Cambridge, MA: MIT Press, 2008.

27. *Teh Y. W., Jordan M. I., Beal M. J., Blei D. M.* Hierarchical Dirichlet processes // Journal of the American Statistical Association. 2004. Vol. 101, N. 476. P. 1566–1581.

28. *Watrous J. D., Alexandrov T., Dorrestein P. C.* The evolving field of imaging mass spectrometry and its impact on future biological research // Journal of Mass Spectrometry. 2011. Vol. 46, N. 2. P. 209–222.

29. *Николенко С. И., Тулупьев А. Л.* Самообучающиеся системы. М.: МЦНМО, 2009. 288 с.

30. *Сироткин А. В.* Байесовские сети доверия: дерево сочленений и его вероятностная семантика // Труды СПИИРАН. 2006. Т. 1, № 3. С. 228–239.

31. *Сироткин А. В.* Вычислительная сложность алгоритмов локального апостериорного вывода в алгебраических байесовских сетях // Труды СПИИРАН. 2011. С. 188–214.

32. *Сироткин А. В., Тулупьев А. Л.* Локальный априорный вывод в алгебраических байесовских сетях: комплекс основных алгоритмов // Труды СПИИРАН. 2007. № 5. С. 100–111.

33. *Тулупьев А. Л., Николенко С. И., Никитин Д. А., Сироткин А. В.* Синтез апостериорных оценок истинности суждений в интегрированных базах знаний: детерминированный вариант // Известия высших учебных заведений: Приборостроение. 2006. № 11. С. 35–39.

34. *Тулупьев А. Л., Николенко С. И., Сироткин А. В.* Байесовские сети: логико-вероятностный подход. СПб.: Наука, 2006. 608 с.

35. *Тулупьев А. Л., Николенко С. И., Сироткин А. В.* Синтез апостериорных оценок при поступлении свидетельств с неопределенностью в интегрированную систему знаний о неточных вероятностях // Известия высших учебных заведений: Приборостроение. 2006. № 11. С. 39–44.

36. *Тулупьев А. Л., Сироткин А. В., Николенко С. И.* Синтез согласованных оценок истинности утверждений в интеллектуальных информационных системах // Известия высших учебных заведений: Приборостроение. 2006. № 7. С. 20–26.

37. *Тулупьев А. Л., Сироткин А. В., Николенко С. И.* Байесовские сети доверия: логико-вероятностный вывод в ациклических направленных графах. СПб.: Изд-во С.-Петербургского ун-та, 2009. 400 с.

Николенко Сергей Игоревич — к.ф.-м.н.; научный сотрудник лаборатории математической логики ПОМИ РАН, доцент кафедры математических и информационных технологий, старший научный сотрудник проблемной лаборатории вычислительной биологии СПбАУ НОЦНТ РАН; snikolenko@gmail.com; СПбАУ НОЦНТ РАН, ул. Хлопина, д. 8, корп. 3, г. Санкт-Петербург, 194021, РФ; р.т. +7(812)297-2145, факс +7(812)448-6998.

Sergey I. Nikolenko — Ph.D.; Researcher at Steklov Mathematical Institute, Adjunct Prof. at the Chair of Mathematical and Information Technologies, Senior Researcher at the Laboratory of Algorithmic Biology at St. Petersburg Academic University; snikolenko@gmail.com; St. Petersburg Academic University, ul. Khlopina, d. 8, korp. 3, St. Petersburg, 194021, Russia; office phone +7(812)297-2145, fax +7(812)448-6998.

Александров Фёдор — к.ф.-м.н.; заведующий лабораторией MALDI спектрометрии, Университет Бремена, Германия; theodore@uni-bremen.de; Zentrum für Technomathematik, Universität Bremen, Bibliothekstr. 1, 28359 Bremen, Germany, тел. +49 (0)421 218-63820, факс +49 (0)421 218-98-63820.

Theodore Alexandrov — Ph.D.; Head of MALDI Imaging Lab, Center for Industrial Mathematics, Universität Bremen, Germany; theodore@uni-bremen.de; Zentrum für Technomathematik, Universität Bremen, Bibliothekstr. 1, 28359 Bremen, Germany, tel. +49 (0)421 218-63820, fax +49 (0)421 218-98-63820.

Чернявский Илья Игоревич — студент СПбАУ НОЦНТ РАН; ilya.chernyavsky@gmail.com; СПбАУ НОЦНТ РАН, ул. Хлопина, д. 8, корп. 3, г. Санкт-Петербург, 194021, РФ; р.т. +7(812)297-2145, факс +7(812)448-6998. Научный руководитель – С.И. Николенко.

Ilya I. Chernyavsky — M.Sc. student at St. Petersburg Academic University; ilya.chernyavsky@gmail.com; St. Petersburg Academic University, ul. Khlopina, d. 8, korp. 3, St. Petersburg, 194021, Russia; office phone +7(812)297-2145, fax +7(812)448-6998. Advisor – S.I. Nikolenko.

Поддержка исследований. Работа была поддержана грантом РФФИ 11-01-00760-а, грантом Президента Российской Федерации для молодых кандидатов

наук МК-6628.2012.1, грантом Президента РФ для ведущих научных школ НШ-3229.2012.1, грантом Правительства РФ 11.G34.31.0018 и Федеральной целевой программой «Научные и научно-педагогические кадры инновационной России» за 2009–2013 гг.

Рекомендовано кафедрой МиИТ СПбАУ НОЦНТ РАН, зав. лаб. Омельченко А.В., д.ф.-м.н.

Статья поступила в редакцию 24.04.2012.

РЕФЕРАТ

Чернявский И.И., Александров Ф., Николенко С.И.

Метод сегментации результатов MALDI-спектрометрии на основе графических вероятностных моделей.

Масс-спектрометрия – один из главных современных методов биохимического анализа. К методам масс-спектрометрии относятся, в частности, методы масс-спектрометрического изображения (imaging mass spectrometry); они активно применяются при изучении образцов тканей, разработке лекарств, поиске биомаркеров и для ряда других задач. Задача сегментации результатов MALDI-спектрометрии заключается в определении структурно-однородных областей образца и их визуализации в виде сегментов на изображении; большой практический интерес представляют как непосредственно картины сегментации, так и характерные массы, выраженные в каждом сегменте.

В работе предлагается подход, ранее не применявшийся к сегментации MALDI-данных: вероятностная кластеризация с последующим сглаживанием. Предобработка при помощи кластеризации применяется для того, чтобы разделить многомерные данные (спектры) в каждой точке изображения на похожие подмножества, тем самым, в частности, сократив размерность. Затем применяется метод сглаживания, основанный на ненаправленных графических моделях (марковских полях).

Метод состоит из двух этапов. На первом этапе к спектрам применяется модель LDA (Latent Dirichlet Allocation), и, таким образом, результатом первого этапа являются распределения на темах в каждой точке и распределения на массах для каждой темы. На втором этапе происходит сглаживание картины сегментации с использованием марковской сети. Таким образом, в предложенном методе для каждой точки выбирается её тема с учётом вероятностей тем в данной точке, полученных при помощи LDA, и значений тем в соседних точках; такие вероятностные результаты хорошо приспособлены для дальнейшего анализа данных спектрометрии.

Предложенные в работе алгоритмы реализованы на практике и применены к набору данных MALDI-спектрометрии, полученных со среза мозга крысы; полученные картины сегментации приводятся в работе.

SUMMARY

Chernyavsky I.I., Alexandrov T., Nikolenko S.I.
Segmentation of MALDI imaging results based on graphical models.

Mass spectrometry is one of the primary techniques of modern biochemical analysis. Mass spectrometry methods include, in particular, imaging mass spectrometry approaches; they are often used to study tissue samples, develop new drugs, look for biomarkers, and a number of other problems. The segmentation problem for MALDI spectrometry results is to find the structurally uniform regions in the sample and visualize them as segments on the image; it is also of significant practical interest not only to find segmented pictures but also find characteristic masses expressed in each segment.

In this work we propose an approach that has not been used before for segmenting MALDI data: probabilistic clustering with subsequent smoothing. The preprocessing clustering step is needed to divide high-dimensional data (spectra) in each pixel of the image into similar subsets, thus, in particular, reducing the dimension. Then we use a smoothing technique based on undirected graphical models (Markov fields).

The method consists of two steps. On the first step, we apply the LDA (Latent Dirichlet Allocation) model to the spectra; thus, results of the first step include topic distributions for each pixel and mass distributions for each topic. On the second step, we smooth out the segmentation picture with a Markov random field. Thus, in the proposed approach we select a topic for every pixel based on topic probabilities in this point obtained by LDA and most likely topics in neighboring pixels; such soft probabilistic results are well suited for further analysis of spectrometry data.

The algorithms proposed in this work have been implemented in practice and applied to a set of MALDI imaging data obtained from a rat's brain section; the resulting segmentation pictures are shown in this work.