Байесовская идентификация параметров смеси нормальных распределений[1]

Ю.А. Дубнов, А.В. Булычев

Аннотация. Рассматривается задача восстановления параметров смеси многомерных нормальных распределений, применяющихся в задачах машинного обучения <<без учителя>>. Предложен метод идентификации моделей, базирующийся на байесовском выводе и принципе максимума апостериорного распределения. В работе описан метод поиска максимума многоэкстремальной функции плотности посредством сэмплирования алгоритмом Метрополиса-Гастингса, приведено качественное и количественное сравнение предложенного алгоритма с EM-алгоритмом для максимизации правдоподобия, а также представлены результаты его работы, как на модельных синтетических примерах, так и на реальных данных из коллекции <<fisheriris>>.

Ключевые слова: смесь нормальных распределений, теорема Байеса, алгоритм Метрополиса-Гастингса, классификация.

Введение

Вероятностные модели смесей распределений встречаются в самых разных задачах прикладной математики, статистики и анализа данных. Так, смеси нормальных распределений успешно зарекомендовали себя в кластерном анализе [1, 2] и таких задачах машинного обучения, как распознавание изображений и речи [3, 4], в то время как комбинации более сложных распределений применяются в эконометрике, например, при анализе финансовых рынков [5].

Задача параметрической идентификации смеси заключается в оценке параметров входящих в нее распределений. В случае нормальных распределений, искомые параметры - это математические ожидания и матрицы ковариаций для каждой из компонент, а также весовые коэффициенты самих компонент в составе смеси. Входными данными для задачи идентификации являются наблюдения случайной величины и информация о типе и количестве предполагаемых компонент.

Одним из наиболее распространенных методов решения данной задачи является применение алгоритма максимизации правдоподобия (Expectation Maximization Algorithm - EM-алгоритм) [6]. EM-алгоритм представляет собой итеративную процедуру подбора параметров и максимизации функции правдоподобия. Несмотря на свою распространенность, EM-алгоритм обладает рядом существенных недостатков, таких как сходимость лишь к локальному максимуму правдоподобия и неустойчивость по начальным данным [7]. Существует множество модификаций базового EM-алгоритма,

[1]Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проекты 16-29-12878, 14-07-00837).

направленных на повышение точности и стабильности работы [9]. Кроме того, последнее время набирают популярность альтернативные методы разделения смесей, в том числе и от отечественных разработчиков [10, 11].

Другой подход к решению описанной выше задачи заключается в применении фундаментальной теоремы Байеса и максимизации апостериорного распределения в пространстве параметров модели [12, 13]. Как и в случае с EM-алгоритмом, необходимыми входными данными здесь являются наблюдения и информация о структуре смеси, однако, эффективность применения байесовского оценивания существенным образом зависит от выбора априорного распределения искомых параметров. Единственно верного выбора, по-видимому, не существует, а исследователи по всему миру до сих пор спорят о преимуществах и недостатках различных априорных распределений [14]. В достаточно полном объеме различные техники выбора априорных распределений освещены в [15]. В данной работе авторы прибегают к выбору априорного распределения по методу Зельднера [17], который заключается в составлении наименее информативной функции плотности, обладающей наибольшей энтропией (подробнее разд. 3).

Применение теоремы Байеса позволяет сконструировать в общем виде апостериорное распределение параметров модели, а их наиболее вероятные значения могут быть получены посредством сэмплирования этого распределения. Для сэмплирования апостериорного распределения используется универсальный алгоритм Метрополиса-Гастингса, одним из преимуществ которого является отсутствие влияния <<проклятия размерности>> [19].

В работе приведено детальное описание метода разделения смеси нормальных распределений с помощью Байесовского оцениавния и алгоритма Метрополиса-Гастингса, а также представлены результаты сравнения с EM-алгоритмом. Эксперименты проводились как на искусственных модельных примерах, так и на наборе реальных данных под названием <<fisheriris>> из коллекции Центра машинного обучения и интеллектуальных систем (Center for Machine Learning and Intelligent Systems, University of California) [21].

1. Постановка задачи разделения смеси

Пусть имеется выборка $\boldsymbol{X}$ значений d −мерной случайной величины x^d, плотность распределения которой описывается следующим законом:

$$x^d \propto \sum_{i=1}^{k} w_i \mathcal{N}\left(x^d, \mu_i, \Sigma_i\right), \tag{1}$$

где $\mu_i = \{\mu_i^1, \mu_i^2, \dots, \mu_i^d\}$ - вектор средних значений i-ой компоненты смеси, Σ_i- его ковариационная матрица, k- количество компонент в смеси, а w_i - коэффициенты этих компонент, причем $\sum_{i=1}^{k} w_i = 1$.

При фиксированном векторе параметров

$$\theta = \{\mu_1, \mu_2 \ \dots, \mu_k, \Sigma_1, \Sigma_2, \dots, \Sigma_k, w_1, w_2, \dots, w_{k-1}\}$$

уравнение (1) представляет собой функцию правдоподобия наблюдения случайной величины x^d. Предполагая независимость наблюдений в выборке объемом N, имеем:

$$\mathcal{L}(X|\theta) = \prod_{s=1}^{N} p(x_s^d|\theta) = \prod_{s=1}^{N} \sum_{i=1}^{k} w_i \mathcal{N}(x_s^d, \mu_i, \Sigma_i) \tag{2}$$

Основная цель задачи параметрической идентификации смеси заключается в оценке параметров θ в уравнении (2) по имеющейся выборке наблюдений, а традиционным методом решения является максимизация правдоподобия [6]. Общее количество искомых параметров зависит от размерности задачи d и количества k компонент в смеси:

$$N_\theta = kd + kd^2 + (k-1) = k(d + d^2 + 1) - 1. \tag{3}$$

Так, для одномерной смеси двух нормальных компонент, количество искомых параметров составит $N_\theta = 5$. Таким образом, одномерная задача разделения смеси трансформируется в 5-тимерную задачу поиска оценок параметров этой смеси. Такой быстрый рост числа искомых оценок непременно сказывается на вычислительной сложности применяемых алгоритмов.

Для многомерных задач $(d \geq 2)$ можно сократить общее число искомых параметров, наложив некоторые ограничения на модель, уменьшив тем самым ее универсальность.

Например, в практическом смысле, удобно предполагать диагональный вид матрицы ковариации, что существенно сокращает вектор параметров: для двумерной смеси с тремя нормальными компонентами длина вектора параметров составит $N_\theta = 14$ вместо 20.

2. Байесовский подход

2.1. Теорема Байеса и выбор априорного распределения

Рассмотрим задачу поиска оценок параметров θ с точки зрения теоремы Байеса [12], согласно которой для восстановления апостериорной плотности параметров $p(\theta|X)$ используется функция правдоподобия (2) и некоторая априорная плотность $\pi(\theta)$:

$$p(\theta|X) \propto \mathcal{L}(X|\theta)\pi(\theta) \quad (4)$$

Различают два типа априорных распределений: информативные и неинформативные [14]. Информативные распределения имеют определенную структуру, выбираемую исходя из имеющихся знаний об объекте исследований. Байесовское оценивание в этом случае, фактически, уточняет на основе наблюдений знания, заложенные в априорных распределениях [16]. Другая ситуация с неинформативными априорными распределениями, являющимися, как правило, равномерными или неопределенными в общем виде, но являющимися решением некоторой задачи оптимизации, зачастую - задачи максимизации энтропии [15].

В данной работе будет использоваться комбинированный вариант априорных распределений, предложенный А. Зельнером [17]. Обозначим через $Z(\theta)$ энтропийный интеграл функции правдоподобия:

$$Z(\theta) = -\int \mathcal{L}(X|\theta)\, log\, \mathcal{L}(X|\theta)\, dx. \quad (5)$$

Функция $Z(\theta)$ содержит в себе агрегированную информацию о выборке данных X. Метод Зельнера заключается в выборе априорного распределения $\pi(\theta)$, максимизирующего разницу

$$G = \int Z(\theta)\pi(\theta)\, d\theta - \int \pi(\theta)\, log\, \pi(\theta)\, d\theta \to \max_{\pi(\theta)}. \quad (6)$$

Первое слагаемое представляет собой усредненное по априорному распределению значение функции $Z(\theta)$, а второе - энтропию априорного распределения. Тогда решение задачи (6) одновременно максимизирует информацию, заложенную в наблюдаемой выборке и энтропию априорного распределения, поэтому выбранное таким образом априорное распределение получило название Maximal Data Information Prior (MDIP) [17].

Выбор именно такого априорного распределения для задачи разделения смеси объясняется необходимостью построения наиболее устойчивых оценок даже при малом объеме выборок входных данных, подверженных в практических задачах неструктурированным помехам [18].

Решение оптимизационной задачи (6) может быть получено в общем виде с использованием производных Гато и основной леммы вариационного исчисления, по аналогии с процедурой максимизации энтропии [20]. Опуская подробности математических выкладок, приведем основные моменты решения. Обозначим искомое решение как $\pi^*(\theta)$, представив функцию $\pi(\theta)$ в виде $\pi(\theta) = \pi^*(\theta) + \alpha l(\theta)$, тогда

$$G(\pi(\theta)) = \int Z(\theta)(\pi^*(\theta) + \alpha l(\theta))\, d\theta - \int (\pi^*(\theta) + \alpha l(\theta))\log(\pi^*(\theta) + \alpha l(\theta))\, d\theta \quad (7)$$

Условие стационарности для (7):

$$\left.\frac{\partial G(\pi(\theta))}{\partial \alpha}\right|_{\alpha=0} = \int Z(\theta)l(\theta)d\theta - \int (\log\pi^*(\theta) + 1)l(\theta)\, d\theta = 0$$

или $Z(\theta) - \log\pi^*(\theta) - 1 = 0$,
следовательно,

$$\pi^*(\theta) = \exp(Z(\theta) - 1) \quad (8)$$

Таким образом, значение априорного распределения $\pi^*(\theta)$ может быть вычислено для любой точки θ. Причем размерность интеграла $Z(\theta)$ соответствует размерности исходной смеси d, что существенно меньше размерности вектора параметров.

Согласно (5) выражения (2) и (8) определяют вид апостериорного распределения параметров модели $p(\theta|X)$, причем размерность этого распределения совпадает с длиной вектора параметров θ. Оценки параметров тогда определяются следующей формулой:

$$\hat{\theta} = \arg\max_\theta p(\theta|X). \quad (9)$$

То есть, оценкой параметров по методу максимума апостериорного распределения (MAP) является точка, в которой достигается глобальный максимум функции апостериорной плотности $p(\theta|X)$.

2.2. Сэмплирование апостериорного распределения

Идея поиска MAP посредством сэмплирования заключается в предварительной генерации набора случайных точек с заданным апостериорным распределнием и выборе в качестве искомых оценок именно той точки, в которой реализуется наибольшее значение плотности. Выражение (9) тогда изменится следующим образом:

$$\hat{\theta} = \arg\max_j p(\theta|X), \quad j = \overline{1,T},$$

где T - объем сгенерированной выборки.

Учитывая нетривиальный вид функции апостериорного распределения и высокую его размерность, традиционные методы генерации случайных величин, такие как метод построения обратной функции и метод исключений, оказываются неприменимы. На помощь приходит алгоритм сэмплирования Метрополиса-Гастингса (МГ), не подверженный <<проклятию размерности>> [19].

Алгоритм МГ реализует построение на основе правил цепи Маркова последовательности случайных величин. Для его работы выбирается вспомогательное распределение $q(x^j, x^{(t)})$, основной задачей которого является быстрая генерация точек-кандидатов x^j на попадание в итоговую последовательность. Итерационный процесс алгоритма МГ зависит от текущего состояния $x^{(t)}$ и представляется следующим образом:

• вычисление x^j по вспомогательному распределению $q(x^j, x^{(t)})$;

• вычисление коэффициента α:

$$\alpha = \frac{p(x^j)}{p(x^{(t)})} \frac{q(x^{(t)}|x^j)}{q(x^j|x^{(t)})};$$

• решение о переходе в состояние x^j:

Если $\alpha \geq 1$:

$$x^{(t+1)} = x^j,$$

Если $\alpha < 1$:

$$x^{(t+1)} = \begin{cases} x^j \text{ с вероятностью } \alpha \\ x^{(t)} \text{ с вероятнгостью } 1-\alpha \end{cases}$$

Сгенерированная таким образом последовательность начинается с некоторого начального состояния $x^{(0)}$, а каждая точка в ней зависит лишь от предыдущего состояния. Для получения начального приближения $x^{(0)}$ можно использовать кластеризацию по методу ближайших соседей (k-means) по аналогии с распространенным выбором начальной точки в семействе EM-алгоритмов [7, 22]. С другой стороны, зависимость от начального состояния может быть искусственно снята, отбросив начальную часть точек последовательности. Некоторые задачи дополнительно предполагают выполнение условия независимости сгенерированных сэмплов, в таком случае полученная последовательность точек <<прореживается>>, оставляя, например, каждую десятую или сотую точку.

Основным препятствием для успешного применения алгоритма Метрополиса-Гастингса с целью поиска MAP является отсутствие подходов, как теоретических, так и экспериментальных, для надежного обеспечения полного покрытия исследуемой области в пространстве параметров θ. Поэтому применение алгоритма МГ требует тонкой индивидуальной настройки не только гиперпараметров самого алгоритма, но и использующегося вспомогательного распределения.

3. Сравнение алгоритмов оценивания

3.1. Описание схемы экспериментов

Для подтверждения работоспособности и эффективности применения предложенного метода разделения смеси нормальных распределений была проведена серия из 4 экспериментов, два из которых основаны на искусственных модельных данных, и два - на широко известном наборе данных <<fisheriris>>, использующимся для тестирования алгоритмов кластеризации [21].

Искусственный набор данных формируется базовым генератором случайных величин, распределенных по нормально закону с заданными параметрами. Проверка проводилась как для

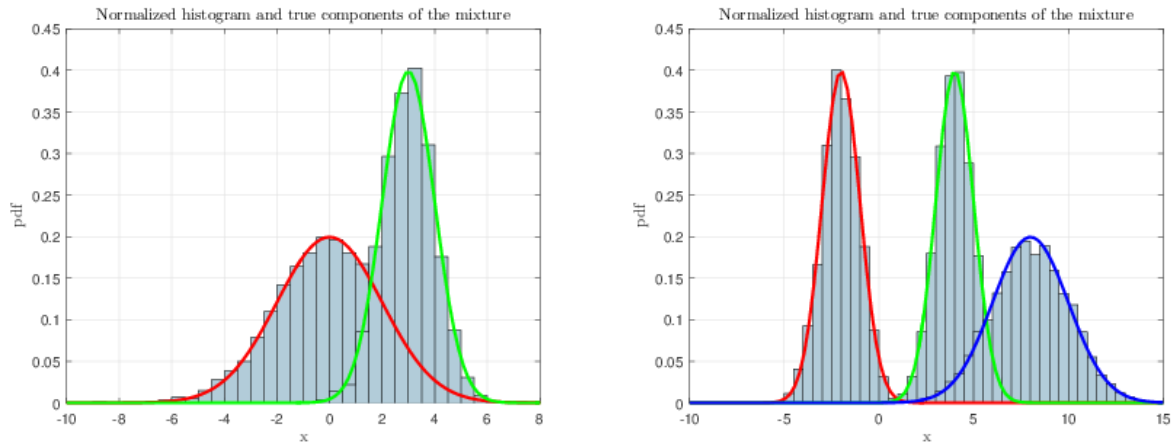

Рис. 1. Нормализованные гистограммы и истинные компоненты смеси для экспериментов 1 (слева) и 2 (справа)

случая малых данных ($N = 50$), так и для случая большого объема входных данных ($N = 1000$). Ключевым этапом при подготовке тестового набора данных является контроль долей вхождения каждой из компонент смеси независимо от ее объема с целью устранения возможного смещения последующих оценок. Что касается готового набора данных <<fisheriris>>, ее объем фиксирован и составляет $N = 150$ примеров, по 50 каждого из 3-х классов.

Приведем кратко параметры тестовых примеров.

1. Двухкомпонентная смесь одномерных нормальных распределений (Рис. 1, слева):

$$x_s \propto \frac{1}{2}\mathcal{N}(0,2) + \frac{1}{2}\mathcal{N}(3,1), \quad s = \overline{1,N},$$

Эксперимент подразумевает оценку параметров двух близких компонент по точкам, которые могут с высокой степенью достоверности (в терминах функции правдоподобия) быть объединены общим нормальным распределением, Длина вектора параметров – $N_\theta = 5$.

2. Трехкомпонентная смесь одномерных нормальных распределений (Рис. 1, справа):

$$x_s \propto \frac{1}{3}\mathcal{N}(-2,1) + \frac{1}{3}\mathcal{N}(4,1) + \frac{1}{3}\mathcal{N}(8,2),$$
$$s = \overline{1,N},$$

Данный эксперимент совпадает с моделью в [22] и нацелен на проверку разделимости смеси с одной явно выраженной отстоящей компонентой. Длина вектора параметров – $N_\theta = 8$.

3. Коллекция <<fisheriris>>, разделение цветков ириса на три группы по данным о длине и ширине лепестков (Рис. 2, слева) с помощью модели трехкомпонентной смеси двумерных нормальных распределений:

$$x_s \propto \sum_{i=1}^{3} w_i \mathcal{N}(\mu_i, \Sigma_i), \quad s = \overline{1,150},$$
$$\mu_i = \{\mu_i^1, \mu_i^2\}, \qquad \Sigma_i = \begin{pmatrix} (\sigma_i^1)^2 & 0 \\ 0 & (\sigma_i^2)^2 \end{pmatrix}.$$

В качестве истинных значений параметров примем выборочные средние значения и стандартные отклонения по кластерам. Длина вектора параметров – $N_\theta = 14$.

4. Коллекция <<fisheriris>>, разделение цветков ириса на три группы по данным размеров как лепестков, так и чашелистиков (Рис. 2, справа) с помощью модели трехкомпонентной смеси четырехмерных нормальных распределений:

$$x_s \propto \sum_{i=1}^{3} w_i \mathcal{N}(\mu_i, \Sigma_i), \quad s = \overline{1,150},$$
$$\mu_i = \{\mu_i^1, \dots, \mu_i^4\}, \qquad \Sigma_i = \begin{pmatrix} (\sigma_i^1)^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (\sigma_i^4)^2 \end{pmatrix}.$$

В этом случае задача усложняется, поскольку 2-ая и 3-я группы имеют схожие параметры чашелистиков. Длина вектора параметров – $N_\theta = 26$.

Для всех экспериментов отдельно оценивается качество полученных оценок по метрике MSE (Mean Square Error) и точность классификации (accuracity). С целью устранения смещения оценок, вызванного процессом генерации

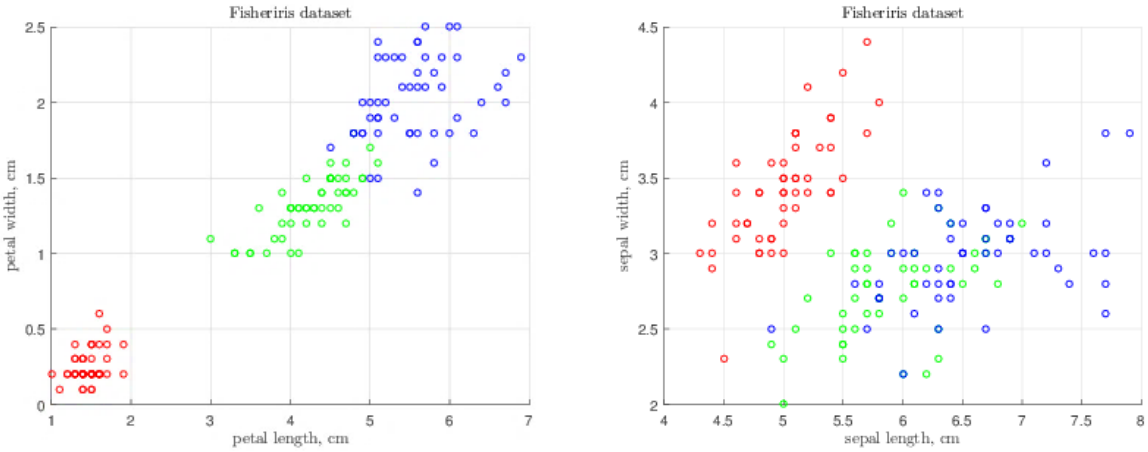

Рис. 2. Распределение примеров коллекции <<fisheriris>> по кластерам в зависимости от размеров лепестков (слева) и чашелистиков (справа)

обучающей выборки, результаты оценивания для модельных экспериментов 1 и 2 усредняются по 100 испытаниям за исключением статистических выбросов.

3.2. Программная реализация

Реализация приведенного в разделе 2 метода восстановления параметров смеси включает 4 последовательных этапа:

1. подготовка набора данных и выбор структуры смеси (фактически, определение размерности поисковой задачи);
2. составление на основе имеющихся наблюдений многокомпонентной функции правдоподобия (2) и вычисление интегральной функции (5);
3. формирование апостериорного распределения по формуле (4) с использованием априорного распределения согласно (8);
4. сэмплирование апостериорного распределения алгоритмом Метрополиса-Гастингса и выбор точки, соответствующей его максимуму.

Расчеты проводились в среде программирования MATLAB версии $2015b$ с использованием инструментов оптимизации (Optimization ToolBox) и статистического обучения (Statistics and Machine Learning Toolbox). В частности, для сэмплирования методом Метрополиса-Гастингса используется функция mhsample, параметры которой рекомендуется подбирать индивидуально для каждой задачи исходя из оптимальности области покрытия. В данном случае в качестве опорного распределения использовались многомерные центрированные нормальные распределения, а объем генерируемой выборки составлял 10^4 точек.

Результатом работы алгоритма являются оценки параметров гауссиан в заданной смеси. Полученные оценки сравниваются с оценками алгоритма максимизации правдоподобия (EM-алгоритм). В качестве прототипа EM-алгоритма использовалось описание, приведенное в монографии Кристофера М. Бишопа [8] (реализация Ali Bahramisharif, 2009, url: www.cs.ru.nl/~ali).

Однако в ряде практических задач, примером которых является эксперимент с данными <<fisheriris>>, гораздо важнее оказывается точность классификации и/или кластеризации, нежели величина ошибки оценок для параметров смеси, причем в таких случаях истинные значения параметров смеси неизвестны, а несмещенность оценок еще не гарантирует высокой точности классификации.

Приведем процедуру определения номера компоненты смеси, породившей ту или иную точку тестового набора. В отличии от EM-алгоритма при использовании сэмплирования и метода MAP результатом являются лишь оценки параметров, поэтому классификация точек проводилась посредством оценки условной вероятности принадлежности точек к каждой из групп, т.е.:

$$c_s = \arg\max_i p(x_s \in C_i|\theta), \quad i = \overline{1,k}, \quad s = \overline{1,N}, \tag{10}$$

где

$$p(x_s \in C_i|\theta) = \frac{w(i)l_i(x_s|\theta)}{\sum_{j=1}^{k} w(j)l_j(x_s|\theta)}, \qquad (11)$$
$$l_i(x_s|\theta) = \mathcal{N}(x_s, \mu_i, \Sigma_i)$$

При истинных значениях параметров θ выражения (10)-(11) представляют собой наивный байесовский классификатор.

3.3. Результаты моделирования

Для демонстрации результатов вектор параметров θ был разбит на 3 группы: матожидания компонент μ_i, диагональные элементы матрицы ковариаций Σ_i и веса компонент w_i для всех компонент смеси $i = 1, \dots, k$. Для каждой группы рассчитывается суммарная среднеквадратическая ошибка (Sum of Mean Squared Errors) по всем измерениям и всем компонентам смеси, например:

$$SMSE_\mu = \sum_{i=1}^{k} MSE(\hat{\mu}_i) = \sum_{i=1}^{k}\sum_{j=1}^{d} MSE(\hat{\mu}_i^j),$$
$$\hat{\mu}_i = \{\hat{\mu}_i^1, \dots, \hat{\mu}_i^d\}.$$

Ошибки оценок для элементов матрицы ковариаций $SMSE_\Sigma$ и для весов компонент $SMSE_w$ рассчитываются аналогичным образом.

Результаты моделирования сведены в Табл. 1 и Табл. 2, где представлены результаты работы EM-алгоритма (EM-estimation) и алгоритма Метрополиса-Гастингса (MH-estimation). В таблице 1 приведены результаты оценок для модельных примеров с синтетическим набором данных разного объема $N = 50,300,1000$, а в Табл. 2 - результаты практических примеров с данными <<fisheriris>>, $N = 150$.

Данные в Табл. 1 наглядно демонстрируют ожидаемое повышение точности обоих методов оценивания с увеличением объема обучающей выборки, в то время как средняя ошибка MH-оценок оказывается меньше в большинстве случаев. Причем, для MH-оценок наибольший скачок точности в 2.38 и 1.13% наблюдается при первичном увеличении выборки от 50 до 300 точек, дальнейшее увеличение обучающей выборки до 1000 точек оказывается не столь эффективным и дает лишь 0.33 и 0.16% для экспериментов 1 и 2 соответственно.

Второй эксперимент отличается большей пространственной разнесенностью компонент смеси, что приводит к высоким показателям точности классификации. Тем не менее, MH-оценивание проигрывает классическому EM-алгоритму при малом объеме входных данных ($N = 50$). Несмотря на меньшую ошибку оценок матожидания и дисперсии, ошибка определения коэффициентов w в этом случае не позволяет достичь высокой точности разделения смеси. Однако оценка разделяющих коэффициентов w значительно улучшается при увеличении объема выборки.

Табл. 1. Результаты экспериментов с модельными данными

Эксперимент # 1, $N_\theta = 5$						
	EM-estimation			*MH-estimation*		
N	50	300	1000	50	300	1000
$SMSE_\mu$	1.2782	0.3396	0.1076	1.3205	0.2753	0.107
$SMSE_\Sigma$	0.5697	0.0931	0.0331	0.5459	0.0874	0.0258
$SMSE_w$	0.0237	0.0113	0.0077	0.285	0.0079	0.0028
acc., %	82.43	84.45	85.17	82.46	84.84	85.17
Эксперимент # 2, $N_\theta = 8$						
	EM-estimation			*MH-estimation*		
N	50	300	1000	50	300	1000
$SMSE_\mu$	1.3005	0.2689	0.1051	1.0944	0.2352	0.1314
$SMSE_\Sigma$	0.6332	0.1402	0.0409	0.6116	0.1236	0.0418
$SMSE_w$	0.004	0.0022	0.002	0.0086	0.0026	0.0009
acc., %	93.74	93.84	93.81	92.78	93.91	94.07

Табл. 2. Результаты экспериментов с реальными данными

Эксперимент # 3, $N_\theta = 14$		
	EM-estimation	*MH-estimation*
	$N = 150$	$N = 150$
$SMSE_\mu$	0.5677	0.008
$SMSE_\Sigma$	0.3788	0.0012
$SMSE_w$	0.0071	0.0016
acc., %	95.33	96.67
Эксперимент # 4, $N_\theta = 26$		
	EM-estimation	*MH-estimation*
	$N = 150$	$N = 150$
$SMSE_\mu$	0.6703	0.1477
$SMSE_\Sigma$	0.4142	0.0394
$SMSE_w$	0.0092	0.0064
acc., %	94.67	96.67

Преимущество использования сэмплирования для определения максимума апостериорного распределения и разделения смеси распределений становится еще более заметным в серии экспериментов с реальными данными (таблица 2), где значительно возрастает размерность пространства параметров N_θ.

В отличии от модельных примеров, при работе с реальными данными из набора <<fisheriris>> ключевым показателем качества оценивания является точность классификации. Использование сэмплирования для разделения смеси позволило увеличить точность классификации по сравнению с результатом EM-алгоритма на 1.34% в случае двумерных данных (эксперимент # 3) и на 2.00% – для 4-мерных данных (эксперимент # 4).

3.4. Влияние зоны покрытия

Увеличение числа компонент смеси и размерности задачи 1 ведет к нелинейному возрастанию количества искомых параметров, в общем случае эта зависимость определяется формулой (3).

Несмотря на то, что алгоритм Метрополиса-Гастингса не подвержен <<проклятию размерностей>> [19], для успешного поиска глобального максимума плотности апостериорного распределения в многомерном пространстве параметров ($N_\theta > 10$) потребуется пропорциональное увеличение количества сэмплов. На Рис. 3 представлена полученная зависимость точности оценивания от количества генерируемых сэмплов для эксперимента # 2 при объеме обучающей выборки $S = 300$ точек.

Из графиков на Рис. 3 видно, что точность определения параметров смеси во многом определяется именно количеством генерируемых сэмплов, что связано, в первую очередь, с многоэкстремальной структурой многомерной функции правдоподобия и, как следствие, плотности апостериорного распределения. Зависимость точности MH-оценок от количества сэмплов во всех рассмотренных примерах оказалась обратно-пропорциональной, и после достижения определенного порога существенное улучшение оценок прекращается.

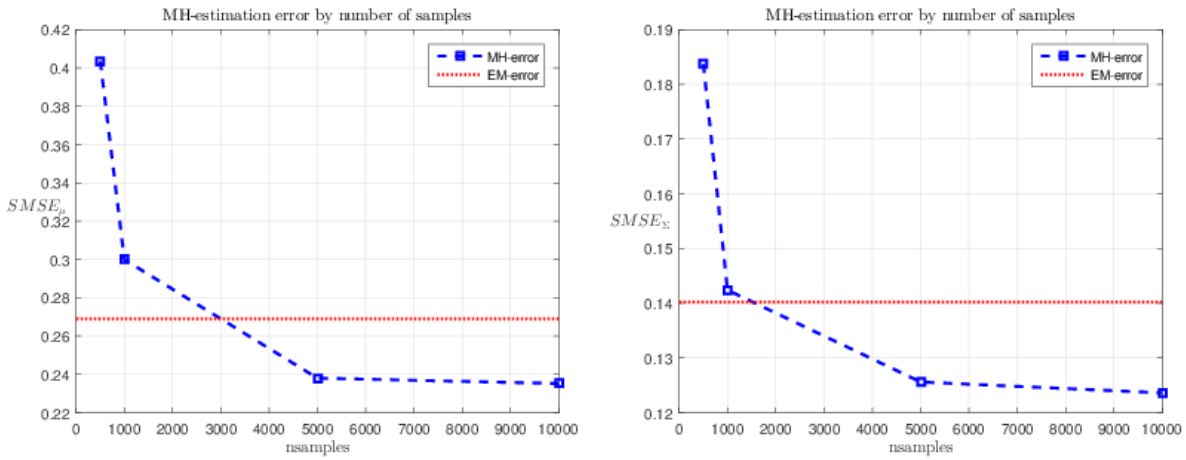

Рис. 3. Зависимость точности MH-оценок параметров матожиданий (слева) и дисперсий (справа) от количества генерируемых сэмплов

Для вычислений использовался ПК с 4-х ядерным процессором Intel(R) Core(TM) i7 CPU 920 @ 2.67 GHz. и 12 Gb оперативной памяти, все полученные в ходе выполнения работы результаты являются строго воспроизводимыми при задании начального состояния генератора случайных чисел (rng(2016)). В плане вычислительных ресурсов наиболее затратным этапом тестирования является сэмплирование случайных векторов, Поэтому, улучшение точности разделения смеси во всех приведенных примерах достигается благодаря использованию значительно больших вычислительных и временных ресурсов по сравнению с применением EM-алгоритма.

Заключение

Рассмотрена задача разделения смеси нормальных распределений. Предложен метод ее решения, базирующийся на поиске максимума байесовской апостериорной плотности посредством сэмплирования алгоритмом Метрополиса-Гастингса. Приведено детальное описание предложенного метода и представлены результаты экспериментов с модельными и реальными данными.

Полученные результаты демонстрируют незначительное но стабильное преимущество применения сэмплирования для определения параметров смеси гауссиан по сравнению с традиционным EM-алгоритмом. Основным недостатком метода максимизации правдоподобия является сходимость к локальному максимуму, поэтому предложенный алгоритм идентификации моделей на основе сэмплирования применим также и для смесей произвольных распределений с гораздо более сложной структурой функции правдоподобия, обладающей множеством локальных экстремумов.

С точки зрения вычислительных ресурсов наиболее трудоемким этапом предложенного метода является генерация ансамбля многомерных случайных векторов, именно поэтому оптимальным решением представляется использование алгоритма Метрополиса-Гастингса, предполагающего многопоточное распараллеливание и реализацию на графических процессорах.

Наиболее актуальной научной проблемой для дальнейшего развития предложенного подхода и успешного применения алгоритма Метрополиса-Гастингса для решения практических задач является настройка его параметров с целью достижения наиболее полной зоны покрытия в многомерном пространстве искомых параметров.

Литература

1. McLachlan G.J. Mixture Models: inference and applications to clustering. – Marcel Dekker, New York, 1988.
2. McLachlan, G., and D. Peel. Finite Mixture Models. – Hoboken, NJ: John Wiley & Sons. Inc., 2000.
3. Figueiredo, M.A.T. and Jain A.K. Unsupervised Learning of Finite Mixture Models. // IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.24(3), pp.381-396, 2012.
4. Reynolds, D.A., Rose, R.C. Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models // IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.3(1). pp.72-83, 1995.
5. Brigo Damiano, Mercurio Fabio. Lognormal-mixture dynamics and calibration to market volatility smiles. // International Journal of Theoretical and Applied Finance, vol.5(4), pp.427-452, 2002.
6. Dempster A.P., Laird N.M., Rubin D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm.// Journal of the Royal Statistical Society. Series B, vol.39(1), pp.1-38, 1977.
7. Xu L., Jordan M.I. On Convergence Properties of the EM Algorithm for Gaussian Mixtures. // Neural Computation, vol.8(1). pp.129-151, 1996.
8. Christopher M. Bishop. Pattern Recognition and Machine Learning. – Springer, 2006 – 758 p.
9. Королев В.Ю. EM-алгоритм, его модификации и применение к задаче разделения смесей вероятностных распределений. – М.: Изд-во ИПИ РАН, 2004. – 102 с. (Korolev V.Yu. EM-algoritm, ego modificacii i primeneniyek zadache razdeleniya smesey veroyatnostnih raspredeleniy. – M.: IPI RAN, 2004. – 102 s.)
10. Королев В.Ю., Назаров А.Л. Разделение смесей вероятностных распределений при помощи сеточных методов моментов и максимального правдоподобия // Автоматика и телемеханика, вып.3, с.98-116, 2010. (Korolev V.Yu., Nazarov A.L. Razdeleniye smesey veroyatnostnih raspredeleniy pri pomoshi setochnih metodov momentov i maksimalnogo pravdopodobiya // Avtomatica i Telemechanika, vip.3, c.98-116, 2010.)
11. Кривенко М.П. Расщепление смеси вероятностных распределений на две составляющие // Информатика и ее применения, т.2, вып.4, с.48-56, 2008. (Krivenko M.P. Rasshepleniye smesi veroyatnostnih raspredeleniy na dve sostavlyaushiye // Informatica i ee primeneniya, t.2, vip.4, s.48-56, 2008.)
12. John E. Rolph. Bayesian Estimation of Mixing Distributions // The Annals of Mathematical Statistics, vol.39, No.4, pp.1289-1302, 1968.
13. Alexander Boulytchev, Vladimir Britkov System modeling of regional economic processes dynamic on the base of the information modeling technology //

Proceedings Of The 10th Eurasia Business And Economics Society Conference (EBES) (May 23-25, 2013, Istanbul, Turkey). – Istanbul: EBES Publications, 2013. – pp.346-354.
14. Andrew Gelman. Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics // Statistical Science, vol.24, No.2, pp.176-178, 2009.
15. Robert E. Kass and Larry Wasserman. The Selection of Prior Distributions by Formal Rules // Journal of the American Statistical Association, vol.91, No.435, pp.1343-1370, 1996.
16. Navid Feroze and Muhammad Aslam. Bayesian Estimation of Two-Component Mixture of Gumbel Type II Distribution under Informative Priors // International Journal of Advanced Science and Technology, vol.53, pp.11-30, 2013.
17. Zellner A. Past and Recent Results on Maximal DataInformation Priors // Texhnical Report, Graduate School of Business, University of Chicago, 1996.
18. Бритков В.Б., Булычев А.В. Информационное моделирование сложных плохоформализуемых систем // Прикладные проблемы управления макросистемами. Под ред. Ю.С. Попкова, В.А. Путилова. – М.: КРАСАНД, 2010. – с.216-231. (Труды Института системного анализа РАН, Т.59) (Britkov V.B., Boulytchev A.V. Informacionnoye modelerovaniye slojnih plohoformalizuemih sistem // Prikladniye problemi upravleniya macrosistemami. Pod redakciyey Yu.S. Popkova, V.A. Putilova. – M.: KRASAND, 2010. – s.216-231. (Trudi Instituta sistemnogo analiza RAN, t.59)).
19. Siddhartha Chib, Edward Greenberg. Understanding the Metropolis-Hastings Algorithm // The American Statistician, vol.49, No.4, pp.327-335, 1995.
20. Popkov Y.S., Dubnov Y.A., Popkov A.Y. New Method of Randomized Forecasting Using Entropy-Robust Estimation: Application to the World Population Prediction. // Mathematics, 2016, vol.4, No.16, pp.1-16.
21. Lichman M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine. CA: University of California, School of Information and Computer Science, 2013.
22. Kamil Dedecius and Jan Reichl. Distributed Estimation of Mixture Models. – Springer International Publishing, 2015. – pp.27-36.

Дубнов Юрий Андреевич. Аспирант МФТИ, математик ИСА РАН (ФИЦ ИУ РАН), преподаватель Высшей школы экономики (ВШЭ). Окончил московский физико-технический институт (государственный университет) в 2013 году. Количество печатных работ: 5. Область научных интересов: динамика макросистем, статистическое обучение, принцип максимума энтропии. E-mail: yury.dubnov@phystech.edu

Булычев Александр Викторович. Ведущий научный сотрудник ИСА РАН (ФИЦ ИУ РАН), Доцент факультета компьютерных наук Высшей школы экономики (ВШЭ). Окончил московский физико-технический институт (государственный университет) в 2006 году. Количество печатных работ: 30. Область научных интересов: анализ данных, байесовские методы в статистике и эконометрике. E-mail: bulytchev.isa.ran@gmail.com

Bayesian Identification of a Gaussian Mixture Model

Yu. A. Dubnov, A. V. Boulytchev

Abstract We consider a problem of parameters estimation for gaussian mixture models widely used in data analysis and unsupervised machine learning. A new model identification method based on Bayesian aproach and the principle of maximum posterior distribution is proposed. In the article we describe the method of multiextremum density function maximum definition using sampling by Metropolis-Hastings algorithm. The proposed method is compared with the traditional expectation maximization algorithm by computational experiments both on a sample synthetic data and the real one from <<fisheriris>> dataset.
Keywords: Gaussian mixture model, Bayesian approach, Metropolis-Hastings algorithm, classification problem.

References

1. McLachlan G.J. Mixture Models: inference and applications to clustering. – Marcel Dekker, New York, 1988.
2. McLachlan, G., and D. Peel. Finite Mixture Models. – Hoboken, NJ: John Wiley & Sons. Inc., 2000.
3. Figueiredo, M.A.T. and Jain A.K. Unsupervised Learning of Finite Mixture Models. // IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.24(3), pp.381-396, 2012.
4. Reynolds, D.A., Rose, R.C. Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models // IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.3(1). pp.72-83, 1995.
5. Brigo Damiano, Mercurio Fabio. Lognormal-mixture dynamics and calibration to market volatility smiles. // International Journal of Theoretical and Applied Finance, vol.5(4), pp.427-452, 2002.
6. Dempster A.P., Laird N.M., Rubin D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm.// Journal of the

Royal Statistical Society. Series B, vol.39(1), pp.1-38, 1977.
7. Xu L., Jordan M.I. On Convergence Properties of the EM Algorithm for Gaussian Mixtures. // Neural Computation, vol.8(1). pp.129-151, 1996.
8. Christopher M. Bishop. Pattern Recognition and Machine Learning. – Springer, 2006 – 758 p.
9. Korolev V.Yu. EM-algoritm, ego modificacii i primeneniyek zadache razdeleniya smesey veroyatnostnih raspredeleniy. – M.: IPI RAN, 2004. – 102 s.
10. Korolev V.Yu., Nazarov A.L. Razdeleniye smesey veroyatnostnih raspredeleniy pri pomoshi setochnih metodov momentov i maksimalnogo pravdopodobiya // Avtomatica i Telemechanika, vip.3, c.98-116, 2010.
11. *Krivenko M.P.* Rasshepleniye smesi veroyatnostnih raspredeleniy na dve sostavlyaushiye // Informatica i ee primeneniya, t.2, vip.4, s.48-56, 2008.John E. Rolph. Bayesian Estimation of Mixing Distributions // The Annals of Mathematical Statistics, vol.39, No.4, pp.1289-1302, 1968.
12. John E. Rolph. Bayesian Estimation of Mixing Distributions // The Annals of Mathematical Statistics, vol.39, No.4, pp.1289-1302, 1968.
13. Alexander Boulytchev, Vladimir Britkov System modeling of regional economic processes dynamic on the base of the information modeling technology // Proceedings Of The 10th Eurasia Business And Economics Society Conference (EBES) (May 23-25, 2013, Istanbul, Turkey). – Istanbul: EBES Publications, 2013. – pp.346-354.
14. Andrew Gelman. Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics // Statistical Science, vol.24, No.2, pp.176-178, 2009.
15. Robert E. Kass and Larry Wasserman. The Selection of Prior Distributions by Formal Rules // Journal of the American Statistical Association, vol.91, No.435, pp.1343-1370, 1996.
16. Navid Feroze and Muhammad Aslam. Bayesian Estimation of Two-Component Mixture of Gumbel Type II Distribution under Informative Priors // International Journal of Advanced Science and Technology, vol.53, pp.11-30, 2013.
17. Zellner A. Past and Recent Results on Maximal DataInformation Priors // Texhnical Report, Graduate School of Business, University of Chicago, 1996.
18. Britkov V.B., Boulytchev A.V. Informacionnoye modelerovaniye slojnih plohoformalizuemih sistem // Prikladniye problemi upravleniya macrosistemami. Pod redakciyey Yu.S. Popkova, V.A. Putilova. – M.: KRASAND, 2010. – s.216-231. (Trudi Instituta sistemnogo analiza RAN, t.59).
19. Siddhartha Chib, Edward Greenberg. Understanding the Metropolis-Hastings Algorithm // The American Statistician, vol.49, No.4, pp.327-335, 1995.
20. Popkov Y.S., Dubnov Y.A., Popkov A.Y. New Method of Randomized Forecasting Using Entropy-Robust Estimation: Application to the World Population Prediction. // Mathematics, 2016, vol.4, No.16, pp.1-16.
21. Lichman M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine. CA: University of California, School of Information and Computer Science, 2013.
22. Kamil Dedecius and Jan Reichl. Distributed Estimation of Mixture Models. – Springer International Publishing, 2015. – pp.27-36.

Dubnov Yury Andreevich. Graduate student of MIPT, mathematician at Institute for Systems Analysis (ISA), FRC CSC RAS, lecturer at Higher School of Economics (HSE). Graduated from Moscow Institute of Physics and Technology (MIPT) in 2013. Author of 5 scientific publications. Research interests: macrosystems dynamic, statistical learning, maximum entropy principle. E-mail: yury.dubnov@phystech.edu

Boulytchev Alexander Viktorovich. Leading Researcher at Institute for Systems Analysis (ISA), FRC CSC RAS, assistant professor at Higher School of Economics (HSE). Graduated from Moscow Institute of Physics and Technology (MIPT) in 2006. Author of 30 scientific publications. Research interests: data analysis, bayesian methods in statistics and econometrics. E-mail: bulytchev.isa.ran@gmail.com