https://www.scilook.eu/index.php/slif/article/view/slif18-01-028
DOI: 10.30888/2415-7538.2020-18-01-028

УДК 519.2

APPLYING THE EM ALGORITHM FOR A GAUSSIAN MIXTURE

ПРИМЕНЕНИЕ АЛГОРИТМА EM ДЛЯ ГАУССОВОЙ СМЕСИ

Pakhomova A.A. / Пахомова А.А.
student / студентка
ORCID: 0000-0001-5104-1134
SPIN: 7013-0127
Li A.D. / Ли А.Д.
student / студентка
ORCID: 0000-0001-6714-9246
Saint Petersburg State University,
Saint Petersburg, Universitetskii prospect 35, 198504
Санкт-Петербургский государственный университет,
Санкт-Петербург, Университетский проспект 35, 198504

Аннотация. *Кластеризация является одной из наиболее важных задач Data Mining. В настоящее время разработано большое количество методов и алгоритмов кластеризации, но, к сожалению, не все они могут эффективно работать с большими массивами данных, поэтому дальнейшие исследования в этом направлении связаны с преодолением этой проблемы. Одним из широко известных в аналитическом сообществе алгоритмов кластеризации, позволяющих эффективно работать с большими объемами данных, является EM-алгоритм. Его название происходит от слов "expectation-maximization", что переводится как "ожидание-максимизация". В статье рассматривается этот эффективный и общий подход, который чаще всего используется для оценки плотности с отсутствующими данными, такие как модель гауссовой смеси.*

Ключевые слова: *анализ данных, кластеризация, алгоритм EM, оценка параметров, оценка максимального правдоподобия, гауссова смесь*

Вступление.

Алгоритм EM был описан и представлен Демпстером [1] в качестве алгоритма получения максимума вероятности, если аналитический расчёт невозможен. Этот алгоритм может быть использован в самых разнообразных ситуациях, когда данные могут рассматриваться как неполные. Помимо очевидных неполных наблюдений, таких как пропущенные данные или цензурированные данные, этот алгоритм может также использоваться в случаях без явных пропущенных значений, как в случае конечных смесей. В таких ситуациях задача состоит в том, чтобы сформулировать проблему как проблему неполных данных для применения алгоритма EM. Широкая сфера применения этого алгоритма в столь многих областях привела к его популярности.

Алгоритм EM [2] имеет много привлекательных преимуществ по сравнению с другими итерационными алгоритмами, такими как алгоритм NR и метод оценки Фишера. К ним относятся, например, экономичность хранения и численная стабильность. Кроме того, в большинстве практических ситуаций этот алгоритм сходится при мягких условиях регулярности к локальному максимуму.

Тем не менее, попытка использовать этот алгоритм во многих статистических ситуациях показывает его ограниченность. Поэтому были разработаны многочисленные модификации и расширения. Особенно тот факт,

что в определенных ситуациях алгоритм сходится очень медленно, привел к разработке модификаций и механизмов ускорения. Например, Реднер и Гомер [3] рекомендовали объединить алгоритм EM с алгоритмом типа Ньютона, где хорошие свойства сходимости алгоритма EM используются вместе с быстрой локальной сходимостью метода Ньютона. Еще один гибридный алгоритм, называемый алгоритмом EM/GN, описан в статье [4], где EM-алгоритм комбинируется с методом Гаусса-Ньютона (GN).

Формулировка алгоритма EM

Ниже приведена формулировка алгоритма EM, как описано у Маклахлан и Кришнан [5]. Каждая итерация алгоритма EM состоит из двух шагов – шага ожидания и шага максимизации, которые кратко называются E- и M-шагами. Эти названия были даны Демпстером [1].

Вкратце, E-шаг содержит подготовку полных данных, которая включает в себя расчет логарифмической вероятности для этого набора. Поскольку это логарифмическое правдоподобие частично основано на ненаблюдаемых данных, оно заменяется его условным ожиданием с учетом наблюдаемых данных с использованием текущего соответствия для неизвестных параметров. Наконец, M-шаг максимизирует это полученное полное логарифмическое правдоподобие по θ. Начиная с подходящих начальных значений параметров, E- и M-шаги затем повторяются, пока не будет выполнен критерий остановки.

Недостающие данные. Пусть $x = (x_1,\ldots,x_n)$ обозначает наблюдаемую выборку размера n, взятую из случайной величины X с функцией плотности вероятности $f(x,\theta)$, где θ – вектор параметров, который должен быть оценен. Пространство параметров обозначается через Ω.

Вводя y в качестве вектора ненаблюдаемых или отсутствующих данных с соответствующей случайной переменной Y, вектор полных данных можно записать как $w = (x,y)$, где W обозначает соответствующую случайную величину. Логарифмическая вероятность для полного набора данных, полная логарифмическая вероятность, обозначается через логарифм $L_c(\theta,w)$.

Expectation-шаг. Пусть $\theta^{(0)}$ – некоторое начальное значение для θ. Затем на первой итерации E-шаг требует вычисления условного математического ожидания полного логарифмического правдоподобия с учетом наблюдаемых данных x с использованием начального значения $\theta^{(0)}$:

$$E_{\theta^{(0)}}[\log L_c(\theta,W)|X = x] =: Q(\theta,\theta^{(0)})$$

Здесь и далее оператор E_θ обозначает ожидание с параметром θ.

Maximization-шаг. На M-шаге Q-функция максимизируется относительно θ в пространстве параметров Ω, что приводит к новой оценке θ, обозначается:

$$\theta^{(1)} = \arg\max_{\theta}(Q(\theta,\theta^{(0)}))$$

Затем E- и M-шаги повторяются, но на этот раз с $\theta^{(1)}$ вместо $\theta^{(0)}$.

На $(t+1)$-ой итерации E- и M-шаги определяются как:

E-step: $Q(\theta,\theta^{(t)}) = E_{\theta^{(t)}}[\log L_c(\theta,W)|X = x]$

M-step: $\theta^{(t+1)} = \arg\max_{\theta}(Q(\theta,\theta^{(t)}))$

Алгоритм EM может быть сформулирован следующим образом:

1. Определение недостающих данных и полных данных;
2. Вычисление условного ожидания полной логарифмической вероятности с учетом наблюдаемых данных, используя некоторую начальную оценку;
3. Максимизация соответствующей Q-функции для получения новой оценки;
4. Замена первоначальной оценки новой оценкой;
5. Повторение шагов 2 и 3 до тех пор, пока не будет достигнут критерий остановки.

Критерий остановки

Е- и М-шаги многократно чередуются до тех пор, пока не будет удовлетворен подходящий критерий остановки. Например, этот процесс обеспечивает последовательность значений наблюдаемого логарифмического правдоподобия. Чтобы остановить итерацию, мы должны рассмотреть абсолютную разницу:

$$\left|\log L\left(\theta^{(t+1)}, x\right) - \log L\left(\theta^{(t)}, x\right)\right|$$

или относительную разницу:

$$\frac{\left|\log L\left(\theta^{(t+1)},\ x\right) - \log L\left(\theta^{(t)},\ x\right)\right|}{\left|\log L\left(\theta^{(t)},\ x\right)\right|}.$$

Если выбранная разница меньше, чем ε — небольшое выбранное значение, алгоритм завершается. Если это происходит на $(t + 1)$-й итерации, то оценка $(t + 1)$ из θ и есть $\hat{\theta} = \theta^{(t+1)}$.

Помимо изменения вероятности, можно также рассмотреть изменение параметров после каждой итерации. На самом деле, в работе [6] показано, что результаты оценки сильно зависят от реализации проекта: «разные стартовые стратегии и правила остановки дают совершенно разные оценки параметра». При этом в литературе окончательно не обсуждается, какой критерий сходимости следует принимать. Хотя наиболее часто используемым критерием остановки является критерий, основанный на вероятности, мы для удобства реализации будем использовать ограниченное число итераций.

Свойства сходимости

После выбора соответствующего критерия остановки интерес представляют свойства сходимости. Они кратко изложены в этом подразделе, а подробное описание можно найти в работах [1, 7], где также показана монотонность логарифмической последовательности правдоподобия. То есть логарифмическая вероятность не уменьшается после итерации EM:

$$\log L\left(\theta^{(t+1)}, x\right) \geq \log L\left(\theta^{(t)}\right)$$

В соответствии с этим выводом и если значения логарифмической вероятности ограничены выше, то последовательность логарифмического правдоподобия почти всегда сходится к некоторому $L^* = L(\theta^*)$.

Поскольку вероятность может иметь несколько стационарных точек, сходимость к максимуму зависит от начальных значений. Однако сближение с локальным или даже с глобальным максимумом не может быть гарантировано. Только в том случае, когда вероятность является унимодальной (и выполняется

условие дифференцирования), любая последовательность EM сходится к единственной оценке, независимо от начальной точки.

Тем не менее, можно найти примеры, когда алгоритм EM сходится к седловой точке, а не к локальному максимуму, как это представлено в работе [5]. Следовательно, нельзя гарантировать, что алгоритм EM сходится к глобальному максимуму, поэтому рекомендация [7] попробовать разные начальные значения представляется целесообразной.

Алгоритм EM для Гауссовых смесей

Уравнение правдоподобия распределения смеси

$$\log L(\Psi, x) = \sum_{s=1}^{n} \log g(x_s, \Psi) = \sum_{s=1}^{n} \log \sum_{j=1}^{k} \pi_j f(x_s, \theta_j)$$

не имеет явного решения. Однако для получения решения может быть применена итерационная процедура.

Предположим, что X — случайная величина с функцией плотности

$$\mathbf{g(x, \Psi) = \sum_{i=1}^{k} \pi_j f(x, \theta_j)}$$

с Гауссовыми компонентами и $\mathbf{\Psi = (\pi_1, \dots, \pi_k, \theta_1, \dots, \theta_k)}$ содержит неизвестные параметры. Выборка $\mathbf{x = (x_1, \dots, x_n)}$ производится на $\mathbf{X}$.

Для оценки параметров смеси может быть применен алгоритм EM со следующими E- и M-шагами:

E-step:

$$Q(\Psi, \Psi^{(t)}) = \sum_{j=1}^{k} \sum_{s=1}^{n} e_{js}^{(t)} \log \pi_j + \sum_{j=1}^{k} \sum_{s=1}^{n} e_{js}^{(t)} \log f(x_s, \theta_j)$$

M-step:

$$\pi_j^{(t+1)} = \frac{1}{n} \sum_{s=1}^{n} e_{js}^{(t)}$$

$$\mu_j^{(t+1)} = \sum_{s=1}^{n} e_{js}^{(t)} x_s \Big/ \sum_{s=1}^{n} e_{js}^{(t)}$$

$$\sigma_j^{2(t+1)} = \sum_{s=1}^{n} e_{js}^{(t)} (x_s - \mu_j^{(t+1)})^2 \Big/ \sum_{s=1}^{n} e_{js}^{(t)},$$

где $e_{js}^{(t)} = \dfrac{\pi_j^{(t)} f(x_s, \theta_t^{(t)})}{g(x_s, \Psi^{(t)})}$

Проблема выбора начальных значений

Некоторые авторы утверждали, что начальные значения не оказывают большого влияния на оценки. Другие подчеркивали необходимость их тщательного рассмотрения, поскольку уравнение правдоподобия часто имеет несколько корней, что может привести к различным решениям.

Для исследования числовых данных существует несколько способов отбора наблюдений. Одни из них представляют нерепрезентативную выборку:

1. «Удобная выборка» — выбираются образцы, которые наиболее удобны. Например, несколько первых значений.
2. «Бессистемная выборка» — отбираем образцы, не думая о способе выбора. Это часто создает иллюзию, что мы выбираем образцы наугад.
3. «Целенаправленная выборка» — отбор проб для конкретной цели. Например, нужно сосредоточиться на крайних случаях. Это может быть полезно, но ограничено, потому что это не позволяет делать заявления о всей совокупности.

Другие же помогают получить репрезентативную выборку:

1. «Простая случайная выборка» — выбор наблюдений (псевдо) случайным образом.

2. «Систематическая выборка» — отбор проб с фиксированным интервалом. Например, каждый 10-й образец (0, 10, 20 и т. д.).

3. «Стратифицированная выборка» — выбор одного и того же количества выборок из разных групп.

4. «Кластерная выборка» — разделение наблюдений на группы (кластеры) и взятие выборки из этих групп.

Таким образом, как определённый метод влияет на наши дальнейшие выводы касательно данных, так и выбор начальных значений влияет на результаты оценок параметров выборки.

Заключение и выводы.

Оценка максимального правдоподобия является сложной задачей для данных при наличии латентных переменных. Но алгоритм EM помогает её итерационно решить. При разных выборах начальных значений метод может давать более точные оценки как с сильно перекрывающимися компонентами, где не очевидно, что распределение состоит из двух субпопуляций, так и хорошо разделенными компонентами, где два пика четко идентифицируются.

Если можно наблюдать два явно хорошо разделенных пика или если существуют примерные знания о двух компонентах, то предложенная процедура приведет к адекватным результатам оценки. Если появляется более двух пиков, представленный метод может быть легко адаптирован.

Тем не менее, иногда гистограмма набора данных, содержащего смеси наблюдений, не выявляет идентифицируемых пиков или неясно, сколько компонентов должно быть установлено. Проблема нахождения правильного числа компонент здесь не исследовалась, однако есть алгоритмы, решающие её.

Хочется отметить, что фундаментальные исследования алгоритма EM были проведены ещё в 20 веке. Сейчас можно найти новые научные труды [8-9], где алгоритм EM подвергается многочисленным модификациям и улучшениям, что показывает его актуальность.

Литература:

1. Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm // Journal of the Royal Statistical Society. Series B

(Methodological), 39(1), pp. 1–38, 1977.

2. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности // (Справочное издание) - М.: Финансы и статистика, 1989. С. 196

3. Redner, R. and Homer, W. F. Mixture densities, maximum likelihood and the EM algorithm // SIAM Review, 26(2), pp. 195–239, 1984.

4. Aitkin, M. and Aitkin, I. A hybrid EM / Gauss-Newton algorithm for maximum likelihood in mixture distributions // Statistics and Computing, 6, pp. 127–130, 1996

5. McLachlan, G. and Krishnan, T. The EM Algorithm and Extensions // Wiley Series in probability and statistics, New York, 1997.

6. Seidel, W., Mosler, K., and Alker, M. A cautionary note on likelihood ratio tests in mixture models // Annals of the Institute of Statistical Mathematics, 52, pp. 481–487, 2000.

7. Wu, C. On the convergence properties of the EM algorithm // The Annals of Statistics, 11(1), pp. 95–103

8. Сташков Д.В., Орлов В.И., Насыров И.Р., Казаковцев Л.А. Применение ЕМ-алгоритма со сферическим гауссовым распределением к задаче классификации промышленной̆ продукции // Экономика и менеджмент систем управления, 2017, Т.23, No1.1, С.185-193. 9. Королев В. Ю. ЕМ-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор // ИПИ РАН. М., 2007

Abstract. *Clustering is one of the most important tasks of Data Mining. Currently, a large number of clustering methods and algorithms have been developed, but, unfortunately, not all of them can work effectively with large data sets, so further research in this direction is related to overcoming this problem. One of the well-known clustering algorithms in the analytical community that allows you to work effectively with large amounts of data is the EM algorithm. Its name comes from the words "expectation-maximization", which translates as "expectation-maximization". This article discusses this efficient and General approach, which is most often used for density estimation with missing data, such as the Gaussian mixture model.*

Key words: *Data Mining, clustering, Expectation-Maximization algorithm, parameter estimation, maximum likelihood estimation, Gaussian mixture*