А.А. ШУМЕЙКО, д.т.н., профессор
А.О. ИСКАНДАРОВА-МАЛА, аспирант
Днепровский государственный технический университет, г. Каменское

О выборе параметров ЕМ-алгоритма для разделения смеси распределений

ЕМ-алгоритм разделения смеси распределений опирается на априорную информацию, как о виде распределения, так и о числе элементов смеси. Предложен метод оценки числа компонентов смеси нормальных распределений. Идея подхода состоит в построении гистограммы с асимптотически оптимальными узлами, восстанавливающей исходные данные с заданной ошибкой.

The EM-algorithm for separation of the mixture of distributions is based on a priori information, both on the type of distribution and on the number of elements in the mixture. A method for estimating the number of components of a mixture of normal distributions is proposed. The idea of the approach is to construct a histogram with asymptotically optimal nodes, recovery the original data with a given error.

Введение

В работе рассмотрен один из видов неиерархической кластеризации — ЕМ-алгоритм [1—3], который традиционно является наиболее используемым инструментом для разделения смеси распределений с известным числом компонентов. Определение числа компонент (кластеров) является нетривиальной задачей. Как правило, этот параметр выбирается исследователем из каких-то априорных предположений [4]. Тем не менее, задача автоматического выбора этого параметра исследовалась во многих работах [5]. Наиболее популярными алгоритмами оценки количества элементов смеси являются:

- метод скользящего контроля,
- принцип минимальной длины описания,
- использование информационного критерия Акаике,
- метод релевантных векторов и пр.

Существующие подходы, как правило, отличаются сложностью реализации, интуитивными подходами или приближенными методами решения сложных аналитических задач, часто не ясно, как они вообще связаны с данной задачей.

ЕМ-алгоритм разделения смеси нормальных распределений, по сути, представляет собой восстановление имеющейся гистограммы линейной комбинацией функций Гаусса, поэтому, совершенно естественно, имея значение заданной погрешности восстановления ε, описать имеющуюся гистограмму обедненной гистограммой со свободными узлами, которая наилучшим образом (то есть с минимальным числом узлов [6]). Полученное число узлов является оценкой числа компонентов смеси, восстанавливаемой линейной комбинацией функций Гаусса с заданной точностью ε. Более того, значения свободных узлов дают возможность получить стартовую оценку остальных параметров смеси.

ЕМ-алгоритм.

Пусть $p(x,\theta_i)$ плотность вероятности того, что наблюдение получено из i-й компоненты смеси (рис. 1) распределений $p(x)=\sum_{i=1}^{k}\omega_i p_i(x)$.

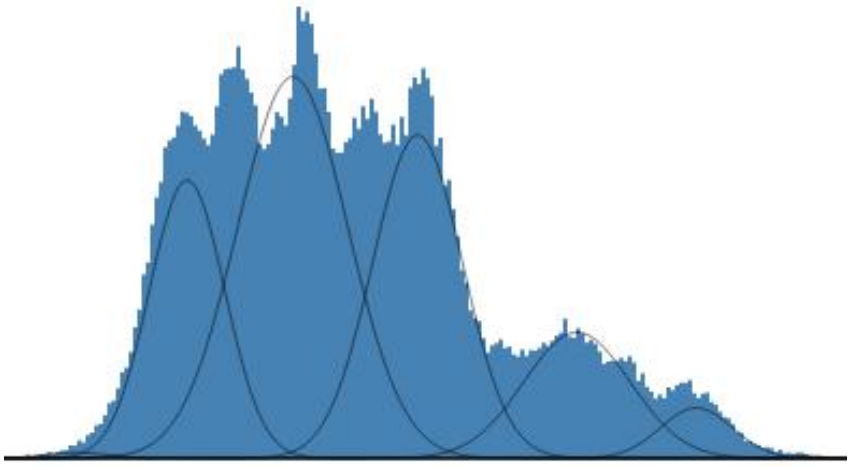

Рис.1. Смесь $p(x)$

Тогда $p(x,\theta_i)=p(x)P(\theta_i|x)=\omega_i p_i(x)$.

Апостериорную вероятность $P(\theta_i|x)$ того, что наблюдение x_j получено из i-й компоненты смеси, обозначим через $g_{i,j}$. Из формулы полной вероятности выпишем условие нормировки

$$\sum_{i=1}^{k} g_{i,j}=1, j=1,...,n.$$

Тогда при известных ω_i и $p_i(x_j)$ из теоремы Байеса легко получить

$$g_{i,j} = \frac{\omega_i p_i(x_j)}{\sum_{\nu=1}^{k} \omega_\nu p_\nu(x_j)}, i = 1,...k, j = 1,...,n.$$

Функция $F(\Theta) = \prod_{j=1}^{n} p(x_j, \theta_i)$ называется функцией правдоподобия от $\Theta = \{(\omega_i, \theta_i)\}_{i=1}^{k}$ по выборке $X = \{x_1,...,x_n\}$.

Традиционно через MLE (Maximum Likelihood Estimate) — метод максимального правдоподобия, называют метод, доставляющий $\hat{\Theta} = \arg\max_{\Theta}(F(\Theta))$.

Заметим, что в силу монотонности алгоритма $\hat{\Theta} = \arg\max_{\Theta}(F(\Theta)) = \arg\max_{\Theta}(\ln(F(\Theta)))$.

Покажем, что при известных $g_{i,j}$, используя MLE, можно получить эффективный метод разделения параметров смеси. Запишем MLE в виде

$$\ln F(\Theta) = \ln \prod_{j=1}^{n} p(x_j, \theta_i) = \sum_{j=1}^{n} \ln \sum_{i=1}^{k} \omega_i p_i(x_j) \to \max_{\Theta}$$

при условии $\sum_{i=1}^{k} \omega_i = 1$.

Используя метод множителей Лагранжа, выпишем лагранжиан этой задачи

$$L(\Theta, \Omega) = \sum_{j=1}^{n} ln \sum_{i=1}^{k} \omega_i p_i(x_j) - \lambda\left(\sum_{i=1}^{k} \omega_i - 1\right).$$

Приравнивая частные производные по неизвестным параметрам нулю, получаем

$$\frac{\partial}{\partial \omega_i} L(\Theta, \Omega) = \sum_{j=1}^{n} \frac{\omega_i p_i(x_j)}{\sum_{\nu=1}^{k} \omega_\nu p_\nu(x_j)} - \lambda = 0, i = 1,...,k..$$

Умножая обе части полученного соотношения на ω_i и суммируя их по всем i, получаем

$$\sum_{j=1}^{n} \sum_{i=1}^{k} \frac{\omega_i p_i(x_j)}{\sum_{\nu=1}^{k} \omega_\nu p_\nu(x_j)} = \lambda . \sum_{i=1}^{k} \omega_i .$$

Замечая, что

$$\sum_{i=1}^{k} \frac{\omega_i p_i(x_j)}{\sum_{\nu=1}^{k} \omega_\nu p_\nu(x_j)} = 1 \text{ и } \sum_{i=1}^{k} \omega_i = 1,$$

Получаем $\sum_{j=1}^{n} 1 = \lambda.$ и $\lambda = n$.

Таким образом имеем

$$\omega_i = \frac{1}{n} \sum_{j=1}^{n} \frac{\omega_i p_i(x_j)}{\sum_{\nu=1}^{k} \omega_\nu p_\nu(x_j)} = \frac{1}{n} \sum_{j=1}^{n} g_{i,j}, i = 1,...,k..$$

Теперь, замечая, что $p_i(x)$ зависит от θ_i, возьмем частную производную лагранжиана от θ_i и приравняем ее нулю

$$\frac{\partial}{\partial \theta_i} L(\Theta, \Omega) = \sum_{j=1}^{n} \frac{\omega_i}{\sum_{\nu=1}^{k} \omega_\nu p_\nu(x_j)} \frac{\partial}{\partial \theta_i} p_i(x_j) = 0, i = 1,...,k.$$

Каждое из слагаемых умножим и разделим на $p_i(x_j)$, соответственно

$$\frac{\partial}{\partial \theta_i} L(\Theta, \Omega) = \sum_{j=1}^{n} \frac{\omega_i p_i(x_j)}{\sum_{\nu=1}^{k} \omega_\nu p_\nu(x_j)} \frac{\partial}{\partial \theta_i} ln\, p_i(x_j) =$$

$$= \sum_{j=1}^{n} g_{i,j} \frac{\partial}{\partial \theta_i} ln\, p_i(x_j) = 0, i = 1,...,k.$$

Отсюда получаем

$$\frac{\partial}{\partial \theta_i} \sum_{j=1}^{n} g_{i,j} \ln p_i(x_j) = 0, i = 1,...,k,$$

что совпадает с необходимым условием максимума в задаче максимизации функции правдоподобия с весом $\sum_{j=1}^{n} g_{i,j}\, ln\, p_i(x_j) \to \max_{\Theta}, i = 1,...,k.$

Таким образом ЕМ-алгоритм с фиксированным числом компонентов смеси можно записать в следующем виде.

Пусть $X = \{x_1,...,x_n\}$ выборка наблюдений, k-число компонентов смеси, $\Theta = \{(\omega_i, \theta_i)\}_{i=1}^{k}$ — начальное приближение параметров смеси, и ε число, определяющее остановку алгоритма.

ЕМ-алгоритм состоит из последовательного применения двух шагов.

Е-шаг (expectation).

$$g_{i,j}^{0} = g_{i,j};\ g_{i,j} = \frac{\omega_i p_i(x_j)}{\sum_{\nu=1}^{k} \omega_\nu p_\nu(x_j)}, i = 1,...k, j = 1,...,n.$$

$$\delta = max\left\{\left|g_{i,j}^{0} - g_{i,j}\right|\right\}$$

М-шаг (maximization).

$$\sum_{j=1}^{n} g_{i,j} \ln p_i(x_j) \to \max_{\Theta}, i = 1,...,k\ ;\ \omega_i = \frac{1}{n} \sum_{j=1}^{n} g_{i,j}, i = 1,...,k.$$

Если $\delta > \varepsilon$, то переходим к Е-шагу, если $\delta \le \varepsilon$, то возвращаем найденные параметры смеси $\Theta = \{(\omega_i, \theta_i)\}_{i=1}^{k}$.

Заметим, что если известен вид функции плотности, то задачу MLE можно выписать в явном виде. Рассмотрим случай, когда известно, что смесь состоит из нормальных распределений $N(\mu_i, \sigma_i^2)$, где $i = 1,...,k$. Тогда задача $\sum_{j=1}^{n} g_{i,j} \ln p_i(x_j) \to \max_{\Theta}, i = 1,...,k$ запишется в виде

$$\sum_{j=1}^{n} g_{i,j} \ln\left(\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x_j - \mu_i)^2}{2\sigma_i^2}\right)\right) \to \max_{\Theta}, i = 1,...,k,$$

или, что то же,

$$G(\{\mu_i, \sigma_i\}_{i=1}^{k}) =$$

$$= \sum_{j=1}^{n} g_{i,j} \left(-\ln(\sigma_i \sqrt{2\pi}) - \frac{(x_j - \mu_i)^2}{2\sigma_i^2}\right) \to \max_{\Theta}, i = 1,...,k.$$

Приравнивая частные производные нулю, получаем

$$\frac{\partial}{\partial \mu_i} G(\{\mu_\nu, \sigma_\nu\}_{\nu=1}^{k}) = -\sum_{j=1}^{n} g_{i,j} \frac{(x_j - \mu_i)}{\sigma_i^2} = 0,$$

отсюда

$$\mu_i = \frac{\sum\limits_{j=1}^{n} g_{i,j} x_j}{\sum\limits_{j=1}^{n} g_{i,j}}.$$

Аналогично,

$$\frac{\partial}{\partial \sigma_i} G\left(\{\mu_\nu, \sigma_\nu\}_{\nu=1}^{k}\right) = -\sum_{j=1}^{n} g_{i,j}\left(\frac{1}{\sigma_i} - \frac{(x_j-\mu_i)^2}{\sigma_i^3}\right) =$$

$$= -\sum_{j=1}^{n} g_{i,j}\left(\frac{\sigma_i^2 - (x_j-\mu_i)^2}{\sigma_i^3}\right) = 0,$$

таким образом, отсюда получаем

$$\sigma_i^2 = \frac{\sum\limits_{j=1}^{n} g_{i,j}(x_j-\mu_i)^2}{\sum\limits_{j=1}^{n} g_{i,j}},$$

что позволяет получить параметры распределения в явном виде при известном числе компонент смеси.

Метод оценки числа компонентов смеси.

Обозначим через $\Delta_n = \{x_i\}_{i=0}^{n}$ — произвольное разбиение $0 = x_0 < x_1 \ldots < x_n = T$, а также положим $h_{i+\frac{1}{2}} = x_{i+1} - x_i$, $h = \max\limits_i h_{i+\frac{1}{2}}$ и $x_{i+\frac{1}{2}} = \frac{x_{i+1}+x_i}{2}$. Через $S\left(\{a_{i+1/2}\}_{i=1}^{n-1}, \Delta_n, x\right)$ обозначим кусочно-постоянную функцию со значениями $a_{i+\frac{1}{2}}$ для $x \in [x_i, x_{i+1})$.

Для функции $y(x)$, непрерывной на [0,*T*], рассмотрим задачу

$$\varepsilon = \min_{x_i} \min_{a_{i+\frac{1}{2}}} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \left(a_{i+\frac{1}{2}} - y(x)\right)^2 dx. \qquad (1)$$

Теорема 1. Пусть $y(x)$ — непрерывно дифференцируемая функция $y \in C^2_{[0,T]}$ такая, что $y'(x)$ на промежутке [0,*T*] имеет не более конечного числа нулей, тогда

$$\varepsilon = \frac{1}{12n^2}\left(\int_0^T \left(y'(x)\right)^{\frac{2}{3}} dx\right)^3 + O\left(\frac{1}{n^3}\right) \qquad (2)$$

и при этом минимум достигается для разбиения $\Delta_n^* = \{x_i^*\}_{i=0}^{m}$ такого, что

$$\int_{x_i^*}^{x_{i+1}^*} \left(y'(x)\right)^{\frac{2}{3}} dx = \frac{1}{n}\int_0^T \left(y'(x)\right)^{\frac{2}{3}} dx \quad \text{и} \quad a^*_{i+\frac{1}{2}} = \frac{1}{h^*_{i+\frac{1}{2}}} \int_{x_i^*}^{x_{i+1}^*} y(x)dx. \qquad (3)$$

Докажем это утверждение. Вначале рассмотрим следующую задачу

$$\Phi\left(a_{i+\frac{1}{2}}\right) = \int_{x_i}^{x_{i+1}} \left(a_{i+\frac{1}{2}} - y(x)\right)^2 dx \to \min_{a_{i+\frac{1}{2}}}. \qquad (4)$$

Тогда ввиду выпуклости задачи (4), необходимое условие минимума является и достаточным, то есть решение этой задачи определяется из уравнения

$$\frac{d\Phi\left(a_{i+\frac{1}{2}}\right)}{da_{i+\frac{1}{2}}} = 2\int_{x_i}^{x_{i+1}} \left(a_{i+\frac{1}{2}} - y(x)\right) dx = 0,$$

и, следовательно,

$$a_{i+\frac{1}{2}} = \frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} y(x)dx$$

Таким образом, задачу (1) можно переписать в виде

$$\varepsilon = \min_{x_i} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \left(y(x) - \frac{1}{h_{i+\frac{1}{2}}}\int_{x_i}^{x_{i+1}} y(x)dx\right)^2 dx.$$

Используя формулу Тейлора

$$y(x) = y_{i+\frac{1}{1}} + y'_{i+\frac{1}{2}}\left(x - x_{i+\frac{1}{2}}\right) + O\left(h^2\right),$$

где $h = \max\limits_i h_{i+\frac{1}{2}}$, получаем

$$\varepsilon = \min_{x_i} \min_{a_{i+\frac{1}{2}}} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \left(a_{i+\frac{1}{2}} - y(x)\right)^2 dx =$$

$$= \min_{x_i} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \left(y_{i+\frac{1}{1}} + y'_{i+\frac{1}{2}}\left(x - x_{i+\frac{1}{2}}\right) + O\left(h^2\right) -\right.$$

$$\left. - \frac{1}{h_{i+\frac{1}{2}}}\int_{x_i}^{x_{i+1}} \left(y_{i+\frac{1}{1}} + y'_{i+\frac{1}{2}}\left(x - x_{i+\frac{1}{2}}\right) + O\left(h^2\right)\right) dx\right)^2 dx =$$

$$= \min_{x_i} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \left(y_{i+\frac{1}{1}} + y'_{i+\frac{1}{2}}\left(x - x_{i+\frac{1}{2}}\right) - y_{i+\frac{1}{1}} + O\left(h^2\right)\right)^2 dx =$$

$$= \min_{x_i} \frac{1}{12}\sum_{i=0}^{n-1}\left(\left(y'_{i+\frac{1}{2}}\right)^2 h^3_{i+\frac{1}{2}} + O\left(h^4\right)\right).$$

Из теоремы о среднем для интегралов следует, что найдется точка $\xi_{i+\frac{1}{2}} \in [x_i, x_{i+1}]$ такая, что

$$\left(y'\left(\xi_{i+\frac{1}{2}}\right)\right)^2 h^3_{i+\frac{1}{2}} = \left(\left(y'\left(\xi_{i+\frac{1}{2}}\right)\right)^{\frac{2}{3}} h_{i+\frac{1}{2}}\right)^3 =$$

$$= \left(\int_{x_i}^{x_{i+1}} \left(y'(x)\right)^{\frac{2}{3}} dx\right)^3,$$

отсюда и из предыдущего сразу получаем

$$\varepsilon = \min_{x_i} \frac{1}{12}\sum_{i=0}^{n-1}\left(\left(y'_{i+\frac{1}{2}}\right)^2 h^3_{i+\frac{1}{2}} + \right.$$

$$\left. + \left(y'\left(\xi_{i+\frac{1}{2}}\right)\right)^2 h^3_{i+\frac{1}{2}} - \left(y'\left(\xi_{i+\frac{1}{2}}\right)\right)^2 h^3_{i+\frac{1}{2}} + O\left(h^4\right)\right) =$$

$$= \min_{x_i} \frac{1}{12} \sum_{i=0}^{n-1} \left(\left(\int_{x_i}^{x_{i+1}} \left(y'(x) \right)^{\frac{2}{3}} dx \right)^3 + O\left(h^4\right) \right). \quad (5)$$

Лемма. Пусть $\alpha > 0$ и $C > 0$, тогда

$$\min \left\{ \sum_{i=0}^{n-1} C_i^{\alpha} \middle| C_i \geq 0, \sum_{i=0}^{n-1} C_i = C \right\} = n \left(\frac{C}{n} \right)^{\alpha},$$

и при этом минимум достигается тогда, когда все C_i равны между собой, то есть,

$$C_i^* = \frac{C}{n} (i = 0,...,n-1).$$

Для доказательства этого утверждения используем метод неопределенных множителей Лагранжа. Выпишем функцию цели

$$L = \lambda_0 \sum_{i=0}^{n-1} C_i^{\alpha} + \lambda_1 \left(\sum_{i=0}^{n-1} C_i - C \right),$$

тогда

$$\frac{\partial L}{\partial C_i} = \lambda_0 \alpha C_i^{\alpha-1} + \lambda_1 = 0,$$

что то же $\lambda_0 \alpha \left(C_i^{\alpha-1} + \lambda_2 \right) = 0$, где $\lambda_2 = \frac{\lambda_1}{\lambda_0 \alpha}$.

Таким образом, получаем $C_i = -\lambda_2^{\frac{1}{\alpha-1}}$ и

$$\sum_{i=0}^{n-1} C_i = -\sum_{i=0}^{n-1} \lambda_2^{\frac{1}{\alpha-1}} = -n \lambda_2^{\frac{1}{\alpha-1}} = C.$$

Следовательно, имеем $\lambda_2^{\frac{1}{\alpha-1}} = -\frac{C}{n}$, или, что то же

$$C_i = \frac{C}{n} (i = 0,...,n-1).$$

Лемма доказана.

Таким образом, из (4) и доказанной леммы получаем

$$\varepsilon = \min_{x_i} \frac{1}{12} \sum_{i=0}^{n-1} \left(\left(\int_{x_i}^{x_{i+1}} \left(y'(x) \right)^{\frac{2}{3}} dx \right)^3 + O\left(h^4\right) \right) =$$

$$= \frac{1}{12} \sum_{i=0}^{n-1} \left(\left(\int_{x_i^*}^{x_{i+1}^*} \left(y'(x) \right)^{\frac{2}{3}} dx \right)^3 + O\left(h^4\right) \right) =$$

$$= \frac{1}{12n^2} \left(\int_0^T \left(y'(x) \right)^{\frac{2}{3}} dx \right)^3 + O\left(\frac{1}{n^3} \right)$$

и при этом минимум достигается для разбиения $\Delta_n^* = \left\{ x_i^* \right\}_{i=0}^m$ такого, что

$$\int_{x_i^*}^{x_{i+1}^*} \left(y'(x) \right)^{\frac{2}{3}} dx = \frac{1}{n} \int_0^T \left(y'(x) \right)^{\frac{2}{3}} dx.$$

Кроме того, заметим, что для заданной погрешности ε можно найти количество n звеньев кусочно-постоянной $S\left(\left\{ a_{i+1/2} \right\}_{i=1}^{n-1}, \Delta_n, x \right)$ с асимптотически оптимальными узлами, выбираемыми из условия (2)

$$n = \left[\sqrt{\frac{1}{12\varepsilon} \left(\int_0^T \left(y'(x) \right)^{\frac{2}{3}} dx \right)^3} \right] + 1,$$

где $[m]$ — целая часть числа m.

Полученное значение количества узлов $S\left(\left\{ a_{i+1/2} \right\}_{i=1}^{n-1}, \Delta_n, x \right)$ может служить оценкой для числа элементов смеси нормальных распределений, а значения $x_{i+1/2}^*$ — стартовыми значениями математического ожидания функций Гаусса.

Вывод

В работе предложен метод определения оценки числа элементов смеси нормальных распределений при известной погрешности восстановления. Важным условием использования данного подхода является адекватное условие смеси именно нормальных распределений. Применение полученного алгоритма показало его эффективность.

ЛИТЕРАТУРА

1. McLachlan G.J. The EM algorithm and extensions / G.J.McLachlan, T.Krishnan .– New York: John Wiley & Sons, Inc., 1997 .– 288 с.
2. Королев В.Ю. ЕМ-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. / В.Ю.Королев; Теоретический обзор .– М.: ИПИ РАН, 2007 .– 94 с.
3. Шумейко А.А. Интеллектуальный анализ данных (Введение в Data Mining) / А.А.Шумейко, С.Л.Сотник .– Днепропетровск: Белая Е.А., 2012 .– 212 с. – Режим доступу: http://pzs.dstu.dp.ua/DataMining/bibl/DataMining.pdf
4. Optimal Histograms with Quality Guarantees [Електронний ресурс] / H.Jagadish, N.Koudas, S.Multhukrishan та ін. – Режим доступу: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.36.769&rep=rep1&type=ps
5. Ветров Д.П. Автоматическое определение количества компонент в ЕМ-алгоритме восстановления смеси нормальных распределений / Д.П.Ветров, Д.А.Кропотов, А.А.Осокин // Ж. вычисл. матем. и матем. физ. – 2010 .– №4(50) .– С.770-783.
6. Лигун А.А. Асимптотические методы восстановления кривых / А.А.Лигун, А.А.Шумейко .– К.: Наукова думка, 1997 .– 358 с .– Режим доступу: http://pzs.dstu.dp.ua/Data/LS_book.pdf

пост. 22.09.2017