УДК 519.6

АЛГОРИТМ ДЛЯ СЕРИИ ЗАДАЧ РАЗДЕЛЕНИЯ СМЕСИ РАСПРЕДЕЛЕНИЙ

Д. В. Сташков*, М. Н. Гудыма, Л. А. Казаковцев, И. П. Рожнов, В. И. Орлов

Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
*E-mail: stashkov@ngs.ru

Представлен генетический алгоритм метода жадных эвристик для задач разделения смеси распределений. Новый алгоритм на основе ЕМ-алгоритма позволяет одновременно решать серию таких задач, различающихся только числом распределений. Статистически показано преимущество нового алгоритма по точности результата для таких задач, как выявление однородных партий электрорадиоизделий.

Ключевые слова: алгоритмы кластеризации, электрорадиоизделия, разделение смеси распределений

ALGORITHM FOR SERIES OF MIXTURE DISTRIBUTION SEPARATION PROBLEMS

D. V. Stashkov*, M. N. Gudyma, L. A. Kazakovtsev, I. P. Rozhnov, V. I. Orlov

Reshetnev Siberian State University of Science and Technology
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russian Federation
*E-mail: stashkov@ngs.ru

We propose new genetic algorithm mixture distribution separation based on ideas of the Greedy Heuristic Method. Based on the EM algorithm, this algorithm allows to solve simultaneously series of such problems with only one various parameter (number of distributions in the mixture). We prove statistically the advantage of our new algorithm by accuracy and stability of its result for such problems as separation homogeneous production batches of microelectronic devices.

Keywords: clustering algorithms, electronic components, separation of mixture distribution.

Данные весьма высокой размерности (несколько сотен измерений) встречаются в задаче выделения однородных партий электронных изделий, например, интегральных схем из сборной партии [1]. В такого рода задачах требуется получение не просто приемлемого результата, но очень точного и стабильного при многократных запусках. Например, такие задачи возникают при проверке качества состава (однородности/неоднородности) смеси однотипных микроэлектронных изделий [2] в космической промышленности. Разделение смеси на предполагаемые однородные партии производится на основе анализа данных тестовых испытаний, представленных векторами данных очень большой размерности (сотни измерений) [1; 3].

Простой ЕМ-алгоритм с двумя чередующимися шагами для разделения смеси распределений [4] в случае многомерных данных сильно зависит от начального решения.

Одной из хорошо зарекомендовавших себя стратегий глобального поиска является применение эволюционных (генетических) алгоритмов. Сложности кодирования решений, традиционно представляемых в классических генетических алгоритмах L-битными строками, в алгоритмах метода жадных эвристик [5] решены применением так называемого генетического алгоритма с вещественным алфавитом, в котором «особи» – промежуточные решения задач k-медиан или k-средних – представлены непосредственно множествами точек в пространстве R^d (т. е. непосредственно множествами медиан или центроидов).

Алгоритм с гетерогенной популяцией для задачи разделения смеси распределений.

1. Сгенерировать случайным образом $N_{POP\text{нач}}$ начальных решений, представленных парой множеств распределений и их весовых коэффициентов $\langle D_m, W_m \rangle = \left\langle \left\{ N(\mu_{m,i}^{\langle 0 \rangle}, \sigma_{m,i}^{\langle 0 \rangle 2}) \right\}, \left\{ \alpha_{m,i}^{\langle 0 \rangle} = 1/k \right\} \right\rangle$, $i = \overline{1, k_m}$, $m = \overline{1, N_{POP\text{нач}}}$. Начальные значения среднеквадратичных отклонений устанавливаются равными для всех кластеров и вычисляются для всей выборки: $\sigma_i^{\langle 0 \rangle 2} = \frac{1}{d} \sum_{x \in S} \left\| x - \bar{x} \right\|^2$. Значения $\mu_{m,i}^{\langle 0 \rangle}$ устанавливаются равными координатам случайно выбранных векторов данных. Для каждого из начальных решений запускается ЕМ-алгоритм, полученные значения целевой функции сохраняются в переменных $f_1, \ldots, f_{N_{POP}}$. Присвоить $N_{iter} = 0$.

2. $N_{iter} = N_{iter} + 1$; $N_{POP} = \max\{N_{POP\text{нач}}; \left\lceil \sqrt{1 + N_{iter}} \right\rceil + 2\}$. Если N_{POP} изменилось, то инициализировать особь X_{Npop} аналогично шагу 1. Выбрать случайным образом $k_1, k_2 \in [1, N_{POP}]$, $k_1 \neq k_2$.

3. $\langle D_{new}, W_{new} \rangle = \langle D_{k_1} \cup D_{k_2}, W_{k_1} \cup W_{k_2} \rangle$.

4. Пока $|D_{new}| > p_{\max}$ выполнять: выбрать $j = \arg \max_{i' \in 1, |D_{new}|} L_{i'} \left(D_{new} \setminus \left\{ N(\mu_{i'}, \sigma_{i'}) \right\}, W_{new} \setminus \left\{ \alpha_{i'} \right\} \right)$;

$D_{new} = D_{new} \setminus \left\{ N(\mu_j, \sigma_j) \right\}$, $W_{new} = W_{new} \setminus \left\{ \alpha_j \right\}$. Следующая итерация 4.

Сравнительные результаты серийного алгоритма

Набор данных, число вектор., размерн.	Число класт. k, тип распр., время	Алгоритм	Ср. рез-т (лог. ф-ция пр-подобия)	Ср.кв. откл. результатов
Europe (UCI), $N = 169308$, $d = 2$	40, сфер, 1.5 часа	Новый	–3625694,1*	20,148
		EM	–3625957,3	49,561
		CEM	–3625779,0	25,064
		SEM	–3625740,2	29,064
Тесты ИС 1526ТЛ1, $N = 1234$, $d = 120$	5, сфер., 5 сек.	Новый	3673,671*	44,043
		EM	3598,160	32,160

Примечание:* – лучший результат;

5. Выбрать случайным образом $p_{child} \in \{2, p_{\max}\}$. Если $p_{child} > |D_{new}|$, то $p_{child} = |D_{new}|$;

6. $f_{child,|D_{new}|} = L\left(\langle D_{new}, W_{new}\rangle\right)$;

7. Пока $|D_{new}| > p_{child}$ выполнять: выбрать $j = \arg\max_{j\in 1,|D_{new}|} L\left(D_{new} \setminus \left\{N(\mu_j, \sigma_j)\right\}, W_{new} \setminus \left\{\alpha_j\right\}\right)$;

$D_{new} = D_{new} \setminus \left\{N(\mu_j, \sigma_j)\right\}, W_{new} = W_{new} \setminus \left\{\alpha_j\right\}$ Следующая итерация 7.

8. Пока $|D_{new}| > 2$: Присвоить $f_{child,|D_{new}|} = L(D_{new}, W_{new})$; $k = |D_{new}|$; $f_{k,|D_{new}|} = L(D_{new}, W_{new})$; если $f_{k,|D_{new}|} < F_k^*$, то присвоить $F_k^* = f_{k,|D_{new}|}$; Выполнить шаги 4.1 и 4.2 для D_{new}. Следующая итерация 8.

9. Выбрать $j_3 \in \{1, N_{POP}\}$ с использованием турнирного замещения. Присвоить

$$D_{j_3} = D_{new}; W_{j_3} = W_{new}; f_{j_3,k} = f_{child}.$$

10. Проверить условия останова, перейти к шагу 2.

Было выполнено по 30 попыток запуска каждого из алгоритмов. Фиксировались лучшие результаты, достигнутые в каждой попытке, затем эти результаты были усреднены. Результаты работы EM-алгоритма в режиме мультистарта и его модификаций обозначены EM, CEM, SEM.

Таким образом, с одной стороны, метод жадных эвристик [5] может быть успешно применен для построения эффективных алгоритмов решения задач разделения смеси распределений. При этом сохраняется важное свойство алгоритмов, полученных с применением данного подхода: высокая точность получаемых результатов. Для некоторых практических задач, к примеру, задачи автоматической группировки электрорадиоизделий [1; 3], сформулированные в виде задач разделения смеси гауссовых распределений результатов тестовых испытаний, новый алгоритм в ходе нескольких (не более 10) попыток запуска позволяет найти, вероятно, точный результат задачи или, по крайней мере, результат, который не получается превзойти с применением известных алгоритмов.

Получен новый алгоритм, стабильно превосходящий по точности получаемых результатов известные алгоритмы для некоторых классов задач, позволяющий получить решение сразу для серии задач разделения смеси распределений. В частности, таким классом задач являются задачи разделения смесей сферических и некоррелированных гауссовых распределений в пространствах большой размерности (десяткисотни измерений) с числом векторов данных от сотен до десятков тысяч.

Библиографические ссылки

1. Федосов В. В., Казаковцев Л. А., Масич И. С. Метод нормировки исходных данных испытаний электрорадиоизделий космического применения для алгоритма автоматической группировки // Системы управления и информационные технологии. 2016. Т. 65 (3). С. 92–96.

2. Федосов В. В. Вопросы обеспечения работоспособности электронной компонентной базы в аппаратуре космических аппаратов : учеб. пособие / Сиб. гос. аэрокосмич. ун-т. Красноярск, 2015. 68 с.

3. Kazakovtsev L. A., Antamoshkin A. N., Masich I. S. Fast Deterministic Algorithm for EEE Components Classification // IOP Conf. Series: Materials Science and Engineering. 2015. Vol. 94. article ID 012015, 10 p. DOI: 10.1088/1757-899X/04/1012015.

4. Королев В. Ю. EM-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор. М. : ИПИ РАН. 2007. 94 с.

5. Казаковцев Л. А., Антамошкин А. Н. Метод жадных эвристик для задач размещения // Вестник СибГАУ. 2015. № 2. С. 317–325.

References

1. Fedosov V. V., Kazakovtsev L. A., Masich I. S. [Method of normalization of raw data of spaceship electronic components testings for automatic grouping algorithm]. *Sistemy upravleniya i informatsionnye tekhnologii*. 2016. Vol. 65, iss. 3. P. 92–96. (In Russ.)

2. Fedosov V. V. *Voprosy obespecheniya rabotosposobnosti elektronnoy komponentnoy bazy v apparature kosmicheskikh apparatov: ucheb.posobie.* [Ensuring the operability of the electronic component base in spacecraft equipment: textbook], Krasnoyarsk, 2015. 68 p.

3. Kazakovtsev L. A., Antamoshkin A. N., Masich I. S. Fast Deterministic Algorithm for EEE Components Classification. IOP Conf. Series: Materials Science and Engineering. 2015. Vol. 94. Article ID 012015, 10 P. DOI: 10.1088/1757-899X/04/1012015.

4. Korolev V. Yu. *EM-algoritm, ego modifikatsii i ikh primenenie k zadache razdeleniya smesey veroyatnostnykh raspredeleniy. Teoreticheskiy obzor.* [EM algorithm, its modifications and their application to the problem of mixture probability distribution separation. Theoretical overview]. Moscow, Institute of Informatics Problems of RAS, 2007. 94 p.

5. Kazakovtsev L. A., Antamoshkin A. N. [Greedy Heuristic Method for Location Problems] // *Vestnik SibGAU*. 2015. Issue 2. P. 317–325. (In Russ.)