

**ВЫЯВЛЕНИЕ ОДНОРОДНЫХ ПАРТИЙ ИЗДЕЛИЙ КОСМИЧЕСКОЙ РАДИОЭЛЕКТРОНИКИ
НА ОСНОВЕ РАЗДЕЛЕНИЯ СМЕСИ СФЕРИЧЕСКИХ ГАУССОВЫХ РАСПРЕДЕЛЕНИЙ**

В. И. Орлов, Д. В. Сташков, Л. А. Казаковцев*, И. Р. Насыров, А. Н. Антамошкин

Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: levk@bk.ru

Выявление однородных партий электрорадиоизделий, применяемых в узлах космической электроники, является одной из важных задач на пути повышения качества этих узлов и, как следствие, срока активного существования и надежности космической техники. Повышение качества достигается как за счет более согласованной работы радиоэлементов с идентичными характеристиками, так и за счет повышения качества и достоверности результатов разрушающих тестовых испытаний, для которых появляется возможность гарантированно отбирать элементы из каждой производственной партии. В настоящей статье задача выделения однородных производственных партий изделий по данным тестовых испытаний решена в виде задачи разделения смеси сферических гауссовых распределений с применением ЕМ-алгоритма с жадной агрегативной эвристикой. ЕМ-алгоритм (алгоритм максимизации математического ожидания), являясь эффективным средством разделения смеси распределений, в случае многомерных гауссовых распределений в пространстве очень высокой размерности оказывается практически неработоспособным: при большом объеме данных требуются слишком громоздкие вычисления для перестроения ковариационных матриц на каждой итерации, при малом объеме данных работа алгоритма приводит к выявлению ложных корреляций. Смешанная (сборная) партия электрорадиоизделий космического применения, представленная многомерным набором данных проведенных над ней неразрушающих тестовых испытаний, рассматривается как смесь сферических гауссовых распределений. Показано, что данная модель в совокупности с новыми разработанными алгоритмами позволяет эффективно выделять однородные партии изделий, может быть применена для разделения достаточно больших сборных партий (тысячи единиц), представленных массивом данных большой размерности (до сотен измерений), и позволяет достичь большей точности и стабильности результата в сравнении с многократным использованием ЕМ-алгоритма в режиме случайного мультистарта.

Ключевые слова: надежность ЭРИ, автоматическая группировка, нечеткая кластеризация.

Sibirskii Gosudarstvennyi Aerokosmicheskii Universitet
imeni Akademika M. F. Reshetneva. Vestnik
Vol. 18, No. 1, P. 69–77**DETECTION OF HOMOGENEOUS PRODUCTION BATCHES OF SPACE ELECTRONIC COMPONENTS
BASED ON SEPARATION OF A MIXTURE OF SPHERICAL GAUSSIAN DISTRIBUTIONS**

V. I. Orlov, D. V. Stashkov, L. A. Kazakovtsev*, I. R. Nasyrov, A. N. Antamoshkin

Reshetnev Siberian State Aerospace University
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russian Federation
*E-mail: levk@bk.ru

Separating of homogeneous production batches of the electronic components used in the electronic units of the space systems is one of the most important problems which must be solved for improving quality of such units, their lifetime and reliability of the space systems. The quality of the electronic units is increased due to both more coordinated work of the EEE components which have identical parameters and increase of quality level and the accuracy of the destructive tests due to a new opportunity of guaranteed selecting electronic elements for these destructive tests from each production batch. In this paper, we solve the problem of precipitations of homogeneous batches of industrial products using Gaussian spherical mixture models and the EM algorithm with agglomerative greedy heuristic procedure. The EM (Expectation Maximization) algorithm is an efficient means of splitting a mix of various distributions. However, in case of multi-dimensional Gaussian distributions in a space of very large dimensionality, this algorithm is actually unworkable. In case of large volume of input data, this algorithm demands too complicated calculation for rebuilding its correlation matrices at each iteration. In case of small data volume, algorithm leads to detection of fake correlation in data. In our paper, the shipped lot of the electronic components for space industry is represented by a data set of non-destructive test results which is considered as a mixture of spherical

Gaussian distributions (SGD). It is shown that this algorithm allows to efficiently determine homogeneous products batches which are rather large (thousands units) using of high-dimensional array of data (up to some hundreds dimensions). We show that, using this mathematical model in combination with new algorithms is capable to separate the homogeneous batches of the electronic components efficiently and reach more accuracy and stability of results in comparison with random multiple start of the algorithm.

Keywords: electronic components reliability, clustering, fuzzy clustering.

Введение. ЕМ-алгоритм (максимизация математического ожидания) успешно применяется для статистических задач, связанных с анализом неполных данных, когда некоторые статистические данные отсутствуют, либо для случаев, когда функция правдоподобия имеет вид, не допускающий удобных методов исследования, но допускающий серьезные упрощения при введении дополнительных «ненаблюдаемых» («скрытых») величин [1]. Именно такая постановка задачи (кластеризация многомерных данных нормального распределения со скрытыми данными) используется нами для решения задач разделения ЭРИ по производственным партиям исходного сырья.

Пусть плотность распределения на множестве X имеет вид смеси k распределений (предполагаем, что распределения гауссовы):

$$\rho(x) = \sum_{j=1}^k \alpha_j \rho_j(x), \quad \sum_{j=1}^k \alpha_j = 1, \quad \alpha_j \geq 0,$$

где $\rho_j(x)$ – функция правдоподобия j -й компоненты смеси; α_j – ее априорная вероятность, а функции правдоподобия принадлежат параметрическому семейству распределений $\phi(x; \theta)$ и отличаются только значениями параметра $\rho_j(x) = \phi(x; \theta)$. Задача разделения смеси (нечеткой кластеризации) заключается в том, чтобы, имея выборку X^N случайных и независимых наблюдений из смеси $\rho(x)$, зная число k и функцию ϕ , оценить вектор параметров распределения $\Theta = (\alpha_1, \dots, \alpha_k, \theta_1, \dots, \theta_k)$.

Искусственно вводится вспомогательный вектор скрытых переменных G , который может быть вычислен, если известны значения вектора параметров Θ .

ЕМ-алгоритм состоит из итерационного повторения двух шагов. На Е-шаге вычисляется ожидаемое значение вектора скрытых переменных G по текущему приближению вектора параметров Θ . Обозначим через $\rho(x, \theta_j) = \rho(x)P(\theta_j|x) = \alpha_j \rho_j(x)$ плотность вероятности того, что объект x получен из j -й компоненты смеси. Обозначим

$$g_{ij} \equiv P(\theta_j | x_i) = \alpha_j \rho_j(x_i) / \sum_{s=1}^k \alpha_s \rho_s(x_i)$$

апостериорную вероятность того, что обучающий объект x_i получен из j -й компоненты смеси.

На М-шаге решается задача максимизации логарифма правдоподобия

$$Q(\Theta) = \ln \prod_{i=1}^N \rho(x_i) = \sum_{i=1}^N \ln \sum_{j=1}^k \alpha_j \rho_j(x_i) \rightarrow \max_{\Theta}$$

и находится следующее приближение вектора Θ по текущим значениям векторов G и Θ . М-шаг сводится к вычислению весов компонент α_j и оцениванию параметров θ_j путем решения k независимых оптимизационных задач.

Данный алгоритм не позволяет определять количество k компонентов смеси (количество кластеров), поэтому должна решаться серия задач с различным предполагаемым числом кластеров. Кроме того, ЕМ-алгоритм обладает сильной неустойчивостью по начальным данным [1]. Тем не менее, к его преимуществам можно отнести [2] то, что его можно комбинировать с другими алгоритмами обработки данных, также для его использования не требуется выделения метрик. При применении более распространенной модели k -средних к задаче выделения однородных производственных партий электрорадиоизделий (ЭРИ) проблема выбора метрики или меры расстояния [3] не имеет однозначного решения. На практике приходится использовать сложные специальные способы нормировки данных [4].

ЕМ-алгоритм успешно работает с малыми объемами данных. Некоторые дорогие виды электрорадиоизделий, несущие в космических аппаратах наибольшую функциональную нагрузку, поступают партиями от нескольких штук, что делает применение ЕМ-алгоритма и алгоритмов на его основе весьма перспективным применительно к нашей задаче.

Простые эксперименты с реализацией ЕМ-алгоритма в среде моделирования R для автоматической группировки партий ЭРИ от 50 до 620 штук (рис. 1, 2) показывают [5], что данный алгоритм способен эффективно разделять смесь распределений, получая адекватные результаты, соответствующие реальным производственным партиям ЭРИ, если для выборки можно выделить несколько (не более 10) информативных признаков.

Кроме определения принадлежности каждого ЭРИ (точки на диаграмме) к кластеру, алгоритм дает таблицу с вероятностными характеристиками принадлежности к кластеру и показывает форму кластера, визуализируя возможные корреляции параметров ЭРИ. В то же время, работа алгоритма при использовании многомерных данных, какими являются данные неразрушающих испытаний ЭРИ, приводит к «зависанию» алгоритма, требует значительных вычислительных ресурсов, особенно в случаях, если число измерений приближается к числу группируемых объектов (векторов данных). Выделение информативных признаков – сложная задача. Так, на рис. 2 видно, что по двум признакам очень трудно выделить партии изделий. В то же время, обработка всех 205 признаков позволяет сделать это с весьма высокой точностью.

Создание модификаций ЕМ-алгоритма с повышенной устойчивостью результата откроет новые перспективы в решении задачи автоматической группировки ЭРИ по производственным партиям, в частности, при небольшом объеме входных данных. Данной проблематике и посвящена настоящая статья.

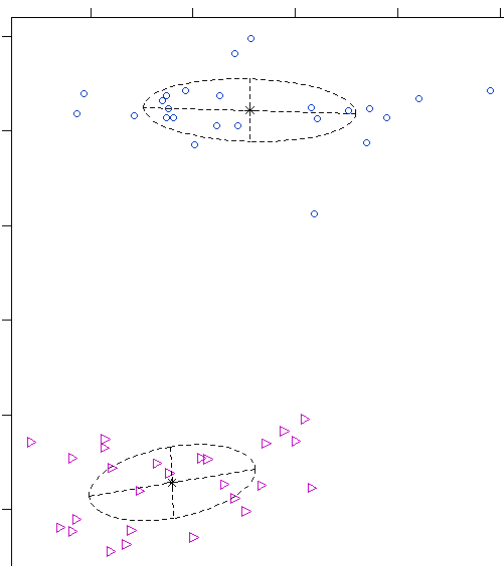


Рис. 1. Срез (показаны 2 параметра из 32) результата автоматической группировки усилителей 140УД25АС1ВК

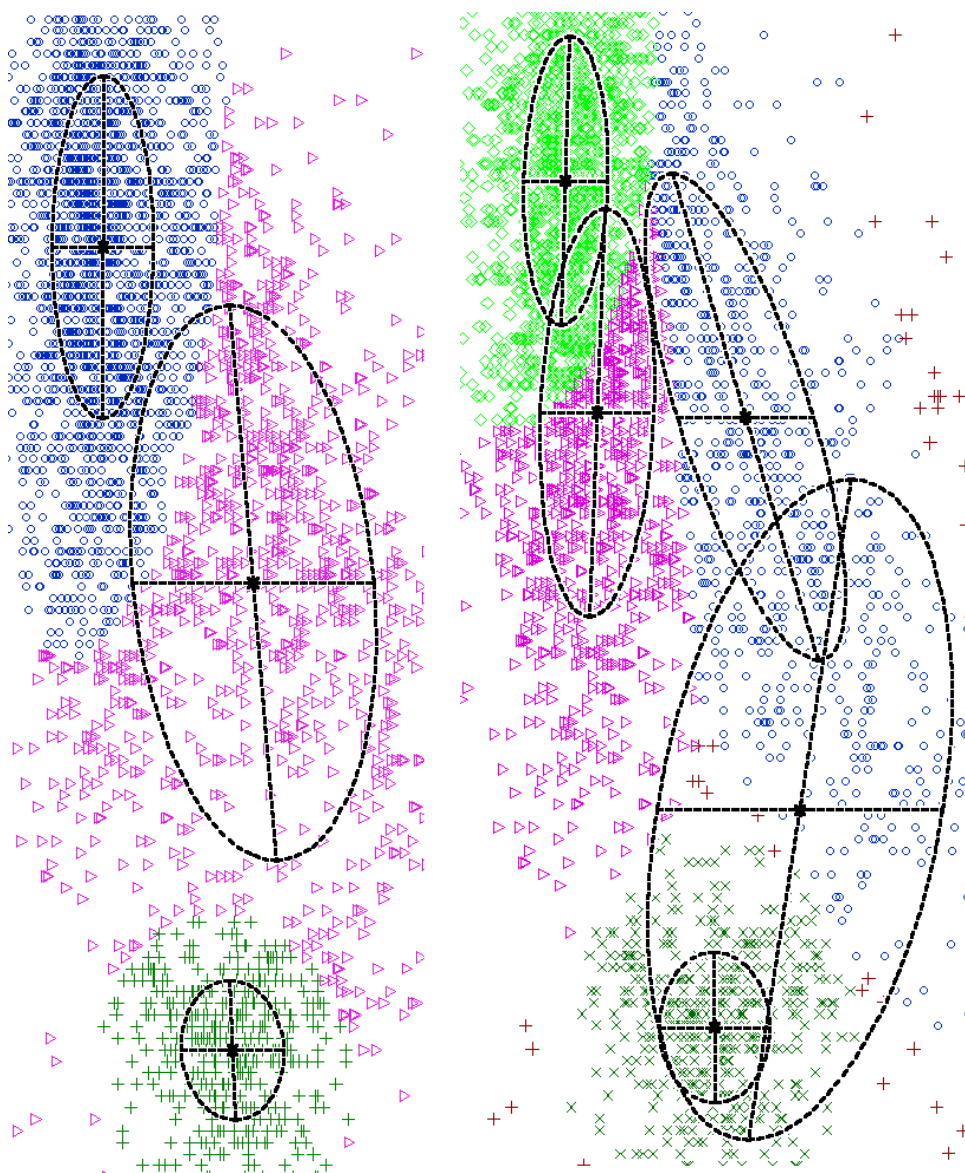


Рис. 2. Разбиение сборной партии ЭРИ 1526ИЕ5 на 3 и 5 групп

Свойства сферических гауссовых распределений. Сферическое гауссово распределение [6] $N(\mu, \sigma^2 I_n)$ определено на $x \in R^n$ с плотностью

$$\rho(x) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left(-\frac{\|x - \mu\|^2}{2\sigma^2}\right),$$

где $\|\cdot\|$ – евклидово расстояние. Если $X = (X_1, \dots, X_n)$ случайно выбраны на распределении $N(0, \sigma^2 I_n)$, тогда их координатами будут независимые случайные переменные с одинаковыми распределениями $N(0, \sigma^2)$. Каждая координата имеет ожидаемое значение дисперсии σ^2 такое, что

$$\mu(\|X\|^2) = \mu(X_1^2 + \dots + X_n^2) = n\sigma^2.$$

Из теории больших отклонений [7], связанной с тем, что $\|X\|^2$ будет плотно сконцентрировано вокруг $n\sigma^2$, следует

$$P(\|X\|^2 - n\sigma^2 > \varepsilon n\sigma^2) \leq e^{-n\varepsilon^2/24}.$$

Таким образом, большая часть плотности вероятности $N(0, \sigma^2 I_n)$ укладывается в тонкую оболочку с радиусом $\sigma\sqrt{n}$ от основания [6]. Это не противоречит факту того, что плотность гауссова распределения будет максимальной у центра, когда площадь поверхности на расстоянии r от основания – $0 \leq r \leq \sigma\sqrt{n}$, увеличивается быстрее, чем уменьшается плотность с расстоянием r [8].

Рассмотрим гауссово распределение $N(0, \sigma^2 I_n)$, имеющее радиус $\sigma\sqrt{n}$. Распределения $N(\mu_1, \sigma_1^2 I_n)$ и $N(\mu_2, \sigma_2^2 I_n)$ на R^n будем называть c -разделенными [9], если выполняется условие $\|\mu_1 - \mu_2\| \geq c \max\{\sigma_1, \sigma_2\} \sqrt{n}$.

Смесь гауссовых распределений является c -разделенной, если все распределения в ней попарно c -разделены. Будем считать, что c_{ij} определяет разделение между i -м и j -м распределением и $c = \min_{i \neq j} c_{ij}$. Мы можем обоснованно ожидать, что сложность исследования смеси гауссовых распределений повышается с понижением значения c .

При $c = 2$ (т. е. при 2-разделенной смеси) кластеры в основном не пересекаются. При больших размерностях пространства n это утверждение верно даже для $\frac{1}{100}$ -разделенной смеси. Если c мало, то проекция

смеси на любую одну координату будет выглядеть унимодальной (частотная гистограмма имеет единственный пик). Это также может быть верно для проекции на несколько координат. Но для больших n , при совместном рассмотрении всех координат, распределение прекратит выглядеть унимодальным (слово «выглядит» обретает буквальный смысл, например, при использовании MDS-визуализации [10]). Именно это обстоятельство часто обуславливает использование данных большой размерности: там, где объекты практически невозможно разделить по одному или нескольким признакам, использование большой совокупности признаков позволяет добиться результата.

Например, в [6] авторы продемонстрировали данное свойство многомерных сферических распределений на 256-мерном наборе дискретных данных, состоящем из рукописных символов, собранных почтовой службой США.

Известные подходы к выделению однородных партий изделий. В настоящей статье задача разделения производственных партий электрорадиоизделий (где каждый кластер должен представлять отдельную производственную партию этих изделий, изготовленную из единой партии сырья) сводится к задаче разделения сферических гауссовых распределений. Задача выделения однородных партий электрорадиоизделий возникает, например, при комплектации узлов космических аппаратов. Специализированные тестовые центры проводят от нескольких десятков до нескольких сотен неразрушающих тестов [11; 12] для каждого экземпляра ЭРИ. При этом предполагается, что все экземпляры ЭРИ уже прошли входной контроль по ужесточенным нормам [13–16], т. е. отдельные их параметры могут колебаться лишь в очень узких диапазонах, определенных этими нормами, и отдельные параметры являются неинформативными. В то же время, совокупность большого числа этих параметров позволяет эффективно разделять производственные партии этих изделий.

В частности, хорошие результаты достигнуты при решении задачи автоматической группировки ЭРИ с использованием моделей k -средних, k -медиан и k -медоид. Недостатком одноименных алгоритмов, основанных на идее процедуры k -средних, называемой также процедурой Ллойда или ALA-процедурой [17; 18], а также РАМ-алгоритма для задачи k -медоид является то, что они относятся к алгоритмам локального поиска (не в строгом смысле), зависящим от выбора начального решения. При этом разработаны детерминированные алгоритмы [19; 20], а также рандомизированные алгоритмы, обеспечивающие, тем не менее, весьма точный и при этом стабильный результат [21–23], что немаловажно во всех процедурах, связанных с космическим производством, все этапы которого требуют весьма строгой регламентации. В частности, хорошие результаты достигаются при применении кластеризации, основанной на расстояниях с прямоугольной метрикой [3].

Преимущество нечеткой кластеризации в виде задачи разделения смеси гауссовых распределений посредством ЕМ-алгоритма заключается в том числе в том, что эта модель, в отличие от модели k -средних и аналогичных, оперирует не абстрактными расстояниями в некоем пространстве характеристик различной физической природы, а вероятностными характеристиками возможности отнесения конкретного экземпляра ЭРИ к той или иной группе (кластеру).

При этом использование именно сферических гауссовых распределений позволяет в некотором смысле обратить так называемое проклятие размерности [24] в полезное свойство многомерных данных. Использование модели традиционных (не сферических) распределений $N(\mu, \Sigma)$ с плотностью распределения

$$\rho(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|} \exp \left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2} \right),$$

где Σ – ковариационная (иначе – корреляционная) матрица, в случае использования данных очень большой размерности приводит к следующим эффектам. Во-первых, матрица размерности n для каждого гауссова распределения (т. е. для каждого кластера) требует достаточно много памяти и значительных вычислительных ресурсов при пересчете ЕМ-алгоритмом на каждом шаге. Во-вторых, использование данной матрицы предполагает, что для каждой координаты каждое распределение (кластер) имеет собственное значение дисперсии (соответствующий элемент диагонали матрицы Σ), при этом плотность j -го распределения тем чувствительнее к изменению координаты, чем меньше дисперсия по этой координате. Таким образом, незначительные колебания какой-либо характеристики, вызванные, например, неточностью измерительных приборов, рассматриваются как значительные при принятии решения об отнесении объекта к тому или иному распределению (кластеру), снижая относительную значимость более информативных признаков.

Тот же эффект проявляется и при кластеризации с использованием модели k -средних в сочетании с нормировкой данных по среднеквадратичному отклонению или 0-1-нормировкой [4]. В то же время, использование специального метода нормировки исходных данных, учитывающего именно физическую природу каждой из характеристик ЭРИ [4], а не разброс характеристики в исследуемой партии изделий, позволяет добиться гораздо лучших результатов.

Аналогичным образом использование единого значения дисперсии по всем координатам (характеристикам), т. е. использование модели смеси сферических гауссовых распределений в сочетании со способом нормировки, описанным в [4], позволяет добиваться адекватных результатов при использовании данных очень большой размерности, которые являются результатом измерений, производимых в ходе тестирования ЭРИ.

Применение ЕМ-алгоритма к задаче выделения однородных партий. ЕМ-алгоритм для разделения смеси сферических распределений может быть описан следующим образом. Дан набор данных $S \subset R^n$, ЕМ-алгоритм для смеси k нормальных распределений с общей сферической матрицей ковариации стартует с начальными стартовыми значениями параметров $\mu_i^{(0)}$, $\alpha_i^{(0)}$, $\sigma_i^{(0)}$, которые в дальнейшем обновляются в соответствии со следующей двухшаговой процедурой (здесь t – номер итерации).

Алгоритм 1. ЕМ-алгоритм.

Шаг 1. Пусть $\tau_i \sim N(\mu_i^{(t)}, \sigma_i^{(t)2} I_n)$ является плотностью i -го гауссова распределения: $\tau_i(x) =$

$$= \frac{1}{(2\pi)^{n/2} \sigma_i^n} \exp \left(-\frac{\|x - \mu_i\|^2}{2\sigma_i^2} \right). \text{ Для каждого вектора}$$

исходных данных $x \in S$ и каждого $1 \leq i \leq k$ по формуле

Байеса [25] вычислим условную вероятность того, что x относится к i -му распределению с учетом текущих параметров:

$$p_i^{(t+1)}(x) = \frac{\alpha_i^{(t)} \tau_i(x)}{\sum_j \alpha_j^{(t)} \tau_j(x)}. \quad (1)$$

Шаг 2. Производится адаптация параметров распределений. Пусть N – количество векторов данных:

$$\alpha_i^{(t+1)} = \frac{1}{N} \sum_{x \in S} p_i^{(t+1)}(x), \quad (2)$$

$$\mu_i^{(t+1)} = \frac{\sum_{x \in S} x p_i^{(t+1)}(x)}{N \alpha_i^{(t+1)}}, \quad (3)$$

$$\sigma_i^{(t+1)2} = \frac{1}{d} \sum_{x \in S} \|x - \mu_i^{(t+1)}\|^2 p_i^{(t+1)}(x) \quad (4)$$

Повторять с шага 1.

Условием останова в таком алгоритме является останов приращения целевой функции, в качестве которой принимается логарифмическая функция правдоподобия

$$L^{(t+1)} = \sum_{x \in S} \sum_{i=1}^k \ln(\tau_i p_i^{(t+1)}(x)). \quad (5)$$

Мы использовали комбинированное условие останова: $t > 100$ или $L^{(t+1)} - L^{(t)} < 0,0001$.

Отметим, что в случае данных большой размерности ($d > 100$) плотность отдельных распределений $\tau_i(x)$ может принимать как очень малые значения ($\tau_i(x) < 1 \cdot 10^{-300}$), так и очень большие значения, что может потребовать применения специальных механизмов при реализации алгоритма. В частности, мы использовали библиотеку decimal [26], применение которой существенно замедляет расчеты, но обеспечивает работу с величинами требуемых порядков.

В работах [21–23] предложен подход к повышению точности и стабильности результата решения задач k -медиан, k -медоид, k -средних, основанный на применении жадных агломеративных эвристических процедур в комбинации с различными метаэвристиками и методами локального поиска. Идея жадной агломеративной эвристики основана на последовательном исключении кластеров из решения. Каждый раз удаляются те кластеры, удаление которых дает наименьший прирост целевой функции (данные задачи являются задачами минимизации).

В настоящей работе мы предлагаем аналогичные подходы к построению жадных агломеративных эвристических процедур. В данном случае мы имеем дело с задачей максимизации функции (5). Первые два подхода полностью аналогичны подходам, примененным в [21–23].

Алгоритм 2. Жадная агломеративная эвристика, вариант 1.

Дано: начальное число гауссовых распределений (кластеров) k , требуемое число кластеров K .

1. Выбрать случайным образом начальное решение с K кластерами, т. е. выбрать случайным образом начальные параметры распределений множества распределений $D = \{N(\mu_i^{(0)}, \sigma_i^{(0)2} I_n), i = \overline{1, k}\}$. В качестве начальных значений $\mu_i^{(0)}$ выбираются случайные векторы данных, дисперсии $\sigma_i^{(0)}$ рассчитываются по формуле (4). Также устанавливаются равные начальные значения множества весовых коэффициентов распределений $W = \{\alpha_i^{(0)}, i = \overline{1, k}\} = \left\{\frac{1}{k}, \dots, \frac{1}{k}\right\}$.

2. Выполнить алгоритм 1, получить новое (улучшенное) решение задачи, представленное множествами D и W .

3. Если $K = k$, то останов.

4. Для каждого $i' \in \{\overline{1, K}\}$ выполнять следующее:

4.1. Получить усеченное множество распределений и множество их весовых коэффициентов $D' = D \setminus \{N(\mu_{i'}^{(0)}, \sigma_{i'}^{(0)2} I_n)\}$, $W' = W \setminus \{\alpha_{i'}^{(0)}\}$.

Запустить алгоритм 1 с начальными значениями параметров распределений, представленных усеченным множеством D' и весами W' . При этом алгоритм 1 ограничивается одной итерацией. Для полученного ЕМ-алгоритмом решения рассчитать целевую функцию L согласно (5), сохранить ее значение в переменной $L_{i'}$.

4.2. Следующая итерация цикла 4.

5. Найти индекс $i'' = \arg \max_{i'=\overline{1, k}} L_{i'}$.

6. Получить усеченные множества

$$D = D \setminus \{N(\mu_{i'}^{(0)}, \sigma_{i'}^{(0)2} I_n)\}, \quad W = W \setminus \{\alpha_{i'}^{(0)}\}.$$

Перейти к шагу 3.

Кроме того, мы испытали два упрощенных алгоритма.

Алгоритм 3. Упрощенная жадная эвристика № 1. Полностью повторяет алгоритм 2, кроме шага 4.2:

4.2. Без запуска ЕМ-алгоритма рассчитать целевую функцию L согласно (5), сохранить ее значение в переменной $L_{i'}$.

Вычислительные эксперименты показывают, что если задать ЕМ-алгоритму в качестве числа требуемых распределений (кластеров) значение k большее, чем фактическое количество кластеров в выборке, во многих случаях в результате работы алгоритма несколько «лишних» кластеров получают значение весовых коэффициентов w_i , близкое к нулю. На этом свойстве основана еще более упрощенная версия алгоритма.

Алгоритм 4. Упрощенная жадная эвристика № 1. Частично повторяет алгоритм 2, при этом цикл 4–4.3 исключен, а шаг 5 выполняется в следующей редакции:

5. Найти индекс $i'' = \arg \max_{i'=\overline{1, k}} L_{i'}$.

Как и в работе о применении жадной агломеративной эвристики с алгоритмами k -средних [27], важным является вопрос выбора начального значения числа кластеров. В настоящей работе мы не рассматривали данный вопрос, ограничиваясь пока экспериментами с $k = 2K$ и $k = 4K$ (удвоенное и учетверенное число начальных кластеров).

Результаты вычислительных экспериментов. Результаты вычислительных экспериментов приведены для двух наборов данных. Первый набор – искусственно сгенерированный набор данных с $m = 200$, $K = 3$, $n = 12$. Второй набор – данные испытаний [28] электрорадиоизделий 1526ИЕ10 ($m = 3987$, $n = 205$), нормированные согласно [4].

Сравнительные результаты работы алгоритмов представлены в табл. 1.

Из табл. 1 видно, что жадная эвристика (алгоритмы 2, 3 и 4) имеет преимущество перед использованием алгоритма 1 по точности результата. В то же время под вопросом остается оптимальное начальное количество кластеров k . Перспективным представляется исследование подхода с адаптацией начального количества кластеров, аналогичный подходу с адаптацией начального количества кластеров для решения задач k -средних и p -медианной задачи, описанному в [27].

Таблица 1

Сравнительные усредненные по 30 запускам результаты работы алгоритмов для двух наборов данных

Набор данных	Алгоритм	Усредненное значение целевой функции (5)	Среднее время работы, с
Сгенерированный набор данных, $m = 200$, $n = 12$, $K = 3$	Алгоритм 1	-7430,39	0,557
	Алгоритм 2, $k = 6$	-7263,48	3,882
	Алгоритм 2, $k = 12$	-7028,99*	21,472
	Алгоритм 3, $k = 6$	-7230,06	3,913
	Алгоритм 3, $k = 12$	-7036,80	23,581
	Алгоритм 4, $k = 6$	-7152,64	2,932
	Алгоритм 4, $k = 12$	-7110,28	9,820
Испытания ЭРИ 1526ИЕ10, $m = 3987$, $n = 205$, $K = 3$	Алгоритм 1	613261,510	14,820
	Алгоритм 2, $k = 6$	622559,536*	113,140
	Алгоритм 2, $k = 12$	622559,536*	653,113
	Алгоритм 3, $k = 6$	622559,536*	96,677
	Алгоритм 3, $k = 12$	622559,536*	667,399
	Алгоритм 4, $k = 6$	622559,536*	96,677
	Алгоритм 4, $k = 12$	613506,5485	553,159

Примечание: * – лучший результат.

**Фрагмент решения задачи автоматической группировки
результатов испытаний ЭРИ 1526ИЕ10, $m = 3987$, $n = 205$, $K = 3$**

№ вектора данных (№ ЭРИ)	Предполагаемый номер кластера	Вероятность отнесения к предполагаемому кластеру
225	1	1
226	1	1
227	1	1
...	...	...
230	2	0,87564
231	1	0,99554
232	1	1
...	...	...

Фрагмент итоговых результатов с указанием оценки вероятности отнесения векторов данных к тому или иному распределению (кластеру) показан в табл. 2. Видно, что в случае многомерных данных вероятности стремятся к 0 или 1, что подтверждает гипотезу о том, что данные являются c -разделенными при малых значениях c .

Заключение. Таким образом, вычислительные эксперименты показывают, что применение жадной агломеративной эвристики позволяет получить результаты значительно более высокой точности в сравнении с мултистартом ЕМ-алгоритма из случайных начальных решений. При этом самая простая и самая быстрая модификация алгоритма с жадной эвристикой – алгоритм 4 – не уступает по точности результатов другим модификациям, требующим гораздо более значительного времени счета. Программное обеспечение на основе описанного подхода может быть использовано в реализации задач повышения надежности программно-аппаратных комплексов автоматизированных систем управления в части управления качеством выпускаемой продукции.

Сходство ЕМ-алгоритма с алгоритмами для решения задач k -средних, k -медоид и k -медиан, а также сходство соответствующих алгоритмов, реализующих жадную агломеративную эвристику, дает обоснованную надежду на получение в дальнейшем алгоритмов, обеспечивающих еще более высокую точность и стабильность результата.

Библиографические ссылки

1. Королев В. Ю. ЕМ-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор / ИПИ РАН. М., 2007. С. 94.
2. Черезов Д. С., Тюкачев Н. А. Обзор основных методов классификации и кластеризации данных // Вестник ВГУ. Сер.: «Системный анализ и информационные технологии». 2009. Вып. № 2. С. 25–29.
3. Казаковцев Л. А., Орлов В. И., Ступина А. А. Выбор метрики при классификации электрорадиоизделий по производственным партиям // Программные продукты и системы. 2015. № 2. С. 124–129.
4. Федосов В. В., Казаковцев Л. А., Гудыма М. Н. Задача нормировки исходных данных испытаний электрорадиоизделий космического применения для алгоритма автоматической группировки // Информационные технологии моделирования и управления. 2016. № 4. С. 263–268.
5. Применение ЕМ-алгоритма к задаче автоматической группировки электрорадиоизделий / В. И. Орлов [и др.] // Решетневские чтения : материалы XX юбилейной Междунар. науч.-практ. конф. (9–12 нояб. 2016, г. Красноярск). Т. 2. С. 72–73.
6. Dasgupta S., Schulman S. J. A Two-Round Variant of EM for Gaussian Mixtures // UAI'00 Proceedings of the Sixteenth Conference on Uncertainty in artificial intelligence. 2000. P. 152–159.
7. Varadhan S. R. S. Special invited paper: Large deviations // The Annals of Probability. 2008. Vol. 36, No. 2. P. 397–419. DOI: 10.1214/07-AOP348.
8. Bishop C. Neural networks for pattern recognition. New York : Oxford University Press, 1995. 498 p.
9. Dasgupta S. Learning mixtures of Gaussians // IEEE Symposium on Foundations of Computer Science. 1999. P. 634–644.
10. Borg J. F. P. Modern Multidimensional Scaling: Theory and Application Springer. 2005. P. 207–212.
11. Федосов В. В., Орлов В. И. Минимально необходимый объем испытаний изделий микроэлектроники на этапе входного контроля // Известия высших учебных заведений. Приборостроение. 2011. Т. 54, № 4. С. 58–62.
12. Задача классификации электронной компонентной базы / Л. А. Казаковцев [и др.] // Вестник СибГАУ. 2014. № 4(56). С. 55–61.
13. Федосов В. В., Патраев В. Е. Повышение надежности радиоэлектронной аппаратуры космических аппаратов при применении электрорадиоизделий, прошедших дополнительные отбраковочные испытания в специализированных испытательных технических центрах // Авиакосмическое приборостроение. 2006. № 10. С. 50–55.
14. Калашников О. А., Никифоров А. Ю. Методика сертификации электронной компонентной базы бортовой космической аппаратуры по стойкости к дозовому воздействию // Спецтехника и связь. 2011. № 4–5. С. 32–38.
15. Калашников О. А., Некрасов П. В., Демидов А. А. Функциональный контроль микропроцессоров при проведении радиационных испытаний // Приборы и техника эксперимента. 2009. № 2. 48 с.
16. Патраев В. Е. Методы обеспечения и оценки надежности космических аппаратов с длительным сроком активного существования : монография / Сиб. гос. аэрокосмич. ун-т. Красноярск, 2010. 136 с.

17. Lloyd S. P. Least Squares Quantization in PCM // *IEEE Transactions on Information Theory*. 1982. Vol. 28. P. 129–137.
18. MacQueen J. B. Some Methods of Classification and Analysis of Multivariate Observations // *Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability*. 1967. Vol. 1. P. 281–297.
19. Казаковцев Л. А. Детерминированный алгоритм для задачи k -средних и k -медоид // *Системы управления и информационные технологии*. 2015. № 1(59). С. 95–99.
20. 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. P. 10. DOI: 10.1088/1757-899X/04/1012015.
21. Казаковцев Л. А., Ступина А. А., Орлов В. И. Модификация генетического алгоритма с жадной эвристикой для непрерывных задач размещения и классификации // *Системы управления и информационные технологии*. 2014. № 2(56). С. 35–39.
22. Modified Genetic Algorithm with Greedy Heuristic for Continuous and Discrete p -Median Problems / L. A. Kazakovtsev [et al.] // *Facta Universitatis (Nis). Series: Mathematics and Informatics*. 2015. Vol. 30, No. 1. P. 89–106.
23. Казаковцев Л. А. Эволюционный алгоритм для задачи k -медоид // *Системы управления и информационные технологии*. 2015. № 2(60). С. 36–40.
24. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning* Springer-Verlag. 2009. 764 P.
25. Гмурман В. Е. *Теория вероятностей и математическая статистика*. М. : Высшее образование, 2005. 400 с.
26. Decimal – Decimal fixed point and floating point arithmetic [Электронный ресурс]. URL: <https://docs.python.org/2/library/decimal.html> (дата обращения: 01.12.2016).
27. Казаковцев Л. А., Антамошкин А. Н. Метод жадных эвристик для задач размещения // *Вестник СибГАУ*. 2015. № 2. С. 317–325.
28. Fuzzy clustering of EEE components for space industry / V. I. Orlov [et al.] // *IOP Conference Series: Materials Science and Engineering*. 2016. Vol. 155. Article ID 012026.
29. *produkty i sistemy*. 2015, No. 2, P. 124–129. Doi: 10.15827/0236-235X.110.124-129 (In Russ.).
4. Fedosov V. V., Kazakovtsev L. A., Gudyma M. N. [Problem of normalization of source testing data of space EEE components clustering algorithm]. *Informatsionnye tekhnologii modelirovaniya i upravleniya*. 2016, No 4, P. 263–268 (In Russ.).
5. Orlov V. I., Stashkov D. V., Gudyma M. N., Kazakovtsev L. A. [EM-algorithm for problem of automatic grouping of electronic components]. *Materialy XX Yubileynoy mezhdunarodnoy nauchno-prakticheskoy konferentsii "Reshetnevskie chteniya"* [Proceed. of XX Anniversary International Scientific and Practical Conference "Reshetnev Readings"]. Krasnoyarsk, 2016, Vol. 2, P. 72–73 (In Russ.).
6. Dasgupta S., Schulman S. J. A Two-Round Variant of EM for Gaussian Mixtures. *UAI'00 Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. 2000, P. 152–159.
7. Varadhan S. R. S. Special invited paper: Large deviations. *The Annals of Probability*. 2008, Vol. 36, No. 2, P. 397–419. Doi:10.1214/07-AOP348.
8. Bishop C. *Neural networks for pattern recognition*. New York : Oxford University Press. 1995, 498 p.
9. Dasgupta S. Learning mixtures of Gaussians. *IEEE Symposium on Foundations of Computer Science*. 1999, P. 634–644.
10. Borg J. F. P. *Modern Multidimensional Scaling: Theory and Application* Springer. 2005, P. 207–212.
11. Fedosov V. V., Orlov V. I. [Minimal necessary extent of examination of microelectronic products at inspection test stage]. *Izvestiya Vuzov. Priborostroenie*. 2011, Vol. 54(4), P. 62–68 (In Russ.).
12. Kazakovtsev L. A., Orlov V. I., Stupina A. A. [Task of electronic components classifying]. *Vestnik SibGAU*. 2014, No. 4(56), P. 55–61 (In Russ.).
13. Fedosov V. V., Patraev V. E. [Increase in reliability of the radio-electronic equipment of spacecrafts at application of the EEE which have passed additional defects tests in the specialized testing technical centers]. *Aviakosmicheskoe priborostroenie*. 2006, No. 10, P. 50–55 (In Russ.).
14. Kalashnikov O. A., Nikiforov A. Yu. [Technique of certification of electronic component base of the onboard space equipment on resistance to dose influence]. *Spetsstekhnika i svyaz'*. 2011, No. 4–5, P. 32–38 (In Russ.).
15. Kalashnikov O. A., Nekrasov P. V., Demidov A. A. [Functional control of microprocessors when carrying out radiation tests]. *Pribory i tekhnika eksperimenta*. 2009, No. 2, 48 p. (In Russ.).
16. Patraev V. E. *Metody obespecheniya i otsenki nadezhnosti kosmicheskikh apparatov s dlitel'nyim srokom aktivnogo sushchestvovaniya* [Methods of providing and assessment of reliability of spacecrafts with the long term of active existence: monograph]. Krasnoyarsk, SibSAU Publ., 2010, 136 p.
17. Lloyd S. P. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*. 1982, Vol. 28, P. 129–137.
18. MacQueen J. B. Some Methods of Classification and Analysis of Multivariate Observations. *Proceedings*

References

1. Korolev V. Ju. *EM-algorithm, ego modifikatsii i ikh primeneniye k zadache razdeleniya smesey veroyatnostnykh raspredeleniy. Teoreticheskiy obzor* [EM algorithm, its modifications and their application to the problem of division of mixes of probabilistic distributions]. Moscow, IPI RAN Publ., 2007, 97 p.
2. Cherezov D. S., Tyukachev N. A. [Review of the main methods of classification and clustering of data]. *Vestnik VGU. Seriya: Sistemnyy analiz i informatsionnye tekhnologii*. 2009, No. 2, P. 25–29 (In Russ.).
3. Kazakovtsev L. A., Orlov V. I., Stupina A. A. [On distance metric for the system of automatic classification of the EEE devices by production batches]. *Programmye*

of the 5th Berkley Symposium on Mathematical Statistics and Probability. 1967, Vol. 1, P. 281–297.

19. Kazakovtsev L. A. [Deterministic algorithm for k -means and k -medoids problems]. *Sistemy upravleniya i informatsionnye tekhnologii*. 2015, No. 1(59), P. 95–99 (In Russ.).

20. Kazakovtsev L. A., Antamoshkin A. N., Masich I. S. Fast deterministic algorithm for EEE components classification problems. *IOP Conference Series: Materials Science and Engineering*. 2015, Vol. 94. Article ID 012015. Doi: 10.1088/1757-899X/94/1/012015.

21. Kazakovtsev L. A., Stupina A. A., Orlov V. I. [Modification of genetic algorithm with greedy heuristics for continuous location and classification problems]. *Sistemy upravleniya i informatsionnye tekhnologii*. 2014, No. 2(56), P. 35–39 (In Russ.).

22. Kazakovtsev L. A., Orlov V. I., Stupina A. A., Kazakovtsev V. L. Modified Genetic Algorithm with Greedy Heuristic for Continuous and Discrete p -Median Problems. *Facta Universitatis (Nis) Series Mathematics and Informatics*. 2015, Vol. 30, No. 1, P. 89–106.

23. Kazakovtsev L. A. [Evolutionary algorithm for k -medoids problem]. *Sistemy upravleniya i informatsionnye tekhnologii*. 2015, No. 2(60), P. 36–40 (In Russ.).

24. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical. *Learning Springer-Verlag*. 2009, P. 764.

25. Gmurman V. E. *Teoriya veroyatnostey i matematicheskaya statistika* [Probability theory and mathematical statistics]. Moscow, Vysshee obrazovanie Publ., 2005, 400 p.

26. Decimal – Decimal fixed point and floating point arithmetic. Available at: URL <https://docs.python.org/2/library/decimal.html> (accessed: 01.12.2016).

27. Kazakovtsev L. A., Antamoshkin A. N. [Greedy heuristic method for location problems]. *Vestnik SibGAU*. 2015, Vol. 15, No. 2, P. 317–325 (In Russ.).

28. Orlov V. I., Stashkov D. V., Kazakovtsev L. A., Stupina A. A. Fuzzy clustering of EEE components for space industry. *IOP Conference Series: Materials Science and Engineering*. 2016, Vol. 155, Article ID 012026.

© Орлов В. И., Сташков Д. В., Казаковцев Л. А., Насыров И. Р., Антамошкин А. Н., 2017