Робастный метод k-means на основе минимизации дифференцируемых оценок среднего, нечувствительных к выбросам

З.М. Шибзухов[1,2], М.А. Казаков[1], Д.П. Димитриченко[1]

[1]Институт математики и информатики МПГУ, пр. Вернадского 88, Москва, Россия, 119991
[2]Институт прикладной математики и автоматизации КБНЦ РАН, ул. Балкарова 2, Нальчик, Россия, 360002

Аннотация. Предложен новый подход к построению варианта алгоритма кластеризации k-means, в котором вместо евклидова расстояния используется расстояние Махаланобиса. Он основан на минимизации дифференцируемых оценок среднего значения, нечувствви-тельных к выбросам. На примерах показана возможность устойчивости предложенного алгоритма по отношению к выбросам в данных.

1. Введение

Классический метод поиска центров и корреляционных матриц кластеров представляет решение следующей задачи минимизации:

$$\mathbf{c}_1^*, \ldots, \mathbf{c}_K^* = \arg \min_{\mathbf{c}_1,\ldots,\mathbf{c}_K} \frac{1}{N} \sum_{k=1}^{N} \min_{j=1,\ldots,K} d(\mathbf{x}_k; \mathbf{c}_j, \mathbf{S}_j), \tag{1}$$

где $\mathbf{c}_1, \ldots, \mathbf{c}_K$ – центры кластеров, $\mathbf{S}_1, \ldots, \mathbf{S}_K$ – корреляционные матрицы,

$$d(\mathbf{x}; \mathbf{c}, \mathbf{S}) = \ln |\mathbf{S}| + (\mathbf{x} - \mathbf{c})' \mathbf{S}^{-1} (\mathbf{x} - \mathbf{c})$$

– квадрат расстояния Махаланобиса с корреляционной матрицей $\mathbf{S}$ между точками $\mathbf{x}$ и $\mathbf{c}$.

Такая постановка задачи основана на предположении о том, точки j-ого кластера подчиняются многомерному нормальному распределению с плотностью

$$p(\mathbf{x}; \mathbf{c}, \mathbf{S}) \propto \frac{1}{\sqrt{|\mathbf{S}|}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c})' \mathbf{S}^{-1} (\mathbf{x}-\mathbf{c})},$$

а произвольная точка $\mathbf{x}$ относится к кластеру с номером

$$j(\mathbf{x}) = \arg \max_{j=1,\ldots,K} p(\mathbf{x}; \mathbf{c}_j, \mathbf{S}_j).$$

Задача (1) сводится к решению систем уравнений:

$$\begin{cases} \mathbf{c}_j = \dfrac{1}{|\mathbf{I}_j|} \displaystyle\sum_{k \in \mathbf{I}_j} \mathbf{x}_k \\ \mathbf{S}_j = \dfrac{1}{|\mathbf{I}_j|} \displaystyle\sum_{k \in \mathbf{I}_j} (\mathbf{x}_k - \mathbf{c}_j)' (\mathbf{x}_k - \mathbf{c}_j), \end{cases} \tag{2}$$

где $\mathbf{I}_j \subset \{1, \ldots, N\}$ – индексы точек, попадающих в j-ый кластер.

Следующая итерационная процедура лежит в основе алгоритма k-means:

$$\begin{cases} \mathbf{c}_{j,t+1} = \dfrac{1}{|\mathbf{I}_{j,t}|} \displaystyle\sum_{k \in \mathbf{I}_{j,t}} \mathbf{x}_k \\ \mathbf{S}_{j,t+1} = \dfrac{1}{|\mathbf{I}_{j,t}|} \displaystyle\sum_{k \in \mathbf{I}_{j,t}} (\mathbf{x}_k - \mathbf{c}_{j,t})'(\mathbf{x}_k - \mathbf{c}_{j,t}), \end{cases} \tag{3}$$

где $\mathbf{I}_{j,t}$ – индексы точек, попадающих в j-ый кластер на t-ом шаге. Начальные значения $\mathbf{c}_{1,0}, \ldots, \mathbf{c}_{K,0}$ и $\mathbf{S}_{1,0}, \ldots, \mathbf{S}_{K,0}$ задаются перед началом итерационной процедуры (3).

Существенное искажение результатов работы алгоритма может появиться, если эмпирическое распределение $\{D(\mathbf{x}_1), \ldots, D(\mathbf{x}_N)\}$, где

$$D(\mathbf{x}) = D(\mathbf{x}; \mathbf{c}_1, \ldots, \mathbf{c}_K; \mathbf{S}_1, \ldots, \mathbf{S}_K) = \min_{j=1,\ldots,K} d(\mathbf{x}; \mathbf{c}_j, \mathbf{S}_j),$$

содержит выбросы.

2. Классический метод преодоления влияния выбросов

Классический метод, направленный на решение проблемы выбросов, основан на замене функции $d(\mathbf{x}; \mathbf{c}, \mathbf{S})$ на

$$d_\varrho(\mathbf{x}; \mathbf{c}, \mathbf{S}) = \ln |\mathbf{S}| + \varrho\big((\mathbf{x} - \mathbf{c})'\mathbf{S}^{-1}(\mathbf{x} - \mathbf{c})\big),$$

где $\varrho(r)$ – функция для подавления влияния выбросов. Она соответствует вероятностному распределению точек с плотностью

$$p(\mathbf{x}; \mathbf{c}, \mathbf{S}) \propto \frac{1}{\sqrt{|\mathbf{S}|}} e^{-\frac{1}{2}\varrho\left((\mathbf{x}-\mathbf{c})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{c})\right)}.$$

Задача оптимизации имеет вид:

$$\mathbf{c}_1^*, \ldots, \mathbf{c}_K^* = \arg \min_{\mathbf{c}_1,\ldots,\mathbf{c}_K} \frac{1}{N} \sum_{k=1}^{N} D_\varrho(\mathbf{x}_k), \tag{4}$$

где

$$D_\varrho(\mathbf{x}) = D_\varrho(\mathbf{x}; \mathbf{c}_1, \ldots, \mathbf{c}_K; \mathbf{S}_1, \ldots, \mathbf{S}_K) = \min_{j=1,\ldots,K} d_\varrho(\mathbf{x}; \mathbf{c}_j, \mathbf{S}_j).$$

Введением функции ϱ можно добиться уменьшения «больших» значений квадрата функции Махаланобиса, если $\varrho(r)$ растет не быстрее, чем линейная. В качестве примера можно привести функцию $\varrho(r) = H(\sqrt{r})$, где H – функция Хьюбера:

$$H(r) = \begin{cases} \frac{1}{2}r^2, & \text{если } r \leqslant c \\ rc - \frac{1}{2}c^2, & \text{если } r < c. \end{cases}$$

Наряду с функцией Хьюбера можно также использовать функцию $S(r) = \sqrt{c^2 + r^2} - c$, которая, в отличие от нее, имеет непрерывную производную второго порядка.

Задача (4) сводится к решению системы уравнений:

$$\begin{cases} \mathbf{c}_j = \frac{1}{V_j} \displaystyle\sum_{k \in \mathbf{I}_j} v_k \mathbf{x}_k, & V_j = \displaystyle\sum_{k \in \mathbf{I}_j} v_k \\ \mathbf{S}_j = \frac{1}{|\mathbf{I}_j|} \displaystyle\sum_{k \in \mathbf{I}_j} v_k (\mathbf{x}_k - \mathbf{c}_j)'(\mathbf{x}_k - \mathbf{c}_j), \end{cases} \tag{5}$$

где $v_k = \psi\big(D_\varrho(\mathbf{x}_k)\big)$, $\psi(r) = \varrho'(r)$.

Для единственности решения необходимо, чтобы $\varrho'(r)$ была неубывающей. Но из этого вытекает, что достаточно сделать выбросами порядка $\frac{1}{n+1}$-ой части набора точек, чтобы сломать робастность такого метода [1]. Тем не менее, если матрицы $\mathbf{S}_1, \ldots, \mathbf{S}_K$ заданы, то задача поиска центров $\mathbf{c}_1, \ldots, \mathbf{c}_K$ сохраняет робастность. Потеря робастности как раз связана с оценкой матриц $\mathbf{S}_1, \ldots, \mathbf{S}_K$.

Достаточно содержательный обзор других методов можно найти в [1,2].

3. Оценки среднего, нечувствительные к выбросам

В настоящей работе предлагается новый подход, основанный на замене среднего арифметического в (1) на робастную дифференцируемую оценку среднего $\mathsf{M}\{z_1, \ldots, z_N\}$, которая будет нечувствительной к выбросам. Такая замена позволит на уровне математической постановки задачи заложить фундамент устойчивости решения задачи.

Такие оценки можно построить, как минимум, двумя способами.

Первый способ основан приближении медианы на базе M-среднего [3,4]

$$\mathsf{M}_\rho\{z_1, \ldots, z_N\} = \arg\min_u \sum_{k=1}^{N} \rho(z_k - u),$$

где ρ – дважды дифференцируемая строго выпуклая функция с минимумом в нуле. Определенное таким образом M-среднее имеет частные производные:

$$\frac{\partial \mathsf{M}_\rho}{\partial z_k} = \frac{\rho''(z_k - \bar{z}_\rho)}{\rho''(z_1 - \bar{z}_\rho) + \cdots + \rho''(z_N - \bar{z}_\rho)},$$

где $\bar{z}_\rho = \mathsf{M}_\rho\{z_1, \ldots, z_N\}$.

Например, если взять функцию $\rho(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon$, то при достаточно малых значениях $\varepsilon > 0$ можно получить приближенный и сглаженный вариант медианы. Выбирая достаточно малое значение ε, можно добиться того, что величина $\rho''(z_k - \bar{z}_\rho) = \varepsilon^2(\varepsilon^2 + (z_k - \bar{z}_\rho)^2)^{-3/2}$ будет пренебрежимо мала для тех значений z_k, которые находятся далеко от среднего значения $\bar{z}_\rho$.

Второй способ основан на применении цензурированного среднего арифметического, в котором пороговое значение оценивается при помощи сглаженного варианта α-квантиля:

$$\mathsf{WM}_{\rho,\alpha}\{z_1, \ldots, z_N\} = \frac{1}{N} \sum_{k=1}^{N} \min\{z_k, \bar{z}_{\rho_\alpha,}\}, \tag{6}$$

где $\rho(r)$ – функция при которой M_ρ выступает в качестве приближения медианы,

$$\rho_\alpha(r) = \begin{cases} \alpha\rho(r), & \text{если } r > 0 \\ \frac{1}{2}\big(\alpha\rho(0_+) + (1-\alpha)\rho(0_+)\big), & \text{если } r = 0 \\ (1-\alpha)\rho(r), & \text{если } r < 0 \end{cases} \tag{7}$$

является функцией, для которой $\mathsf{M}_{\rho_\alpha}\{z_1, \ldots, z_N\}$ выступает в качестве приближения α-квантиля. Частные производные имеют вид:

$$\frac{\partial \mathsf{WM}_{\rho_\alpha}}{\partial z_k} = \begin{cases} \frac{1}{N} + \frac{m}{N}\frac{\partial \mathsf{M}_{\rho\alpha}}{\partial z_k}, & \text{если } z_k < \bar{z}_{\rho_\alpha} \\ \frac{m}{N}\frac{\partial \mathsf{M}_{\rho\alpha}}{\partial z_k}, & \text{если } z_k \geqslant \bar{z}_{\rho_\alpha}, \end{cases}$$

где m – число значений $z_k \geqslant \bar{z}_{\rho_\alpha}$. В обоих случаях $\frac{\partial \mathsf{M}}{\partial z_k} \geqslant 0$ и $\frac{\partial \mathsf{M}}{\partial z_1} + \cdots + \frac{\partial \mathsf{M}}{\partial z_N} = 1$.

Если взять функцию Хьюбера, то можно получить другой вариант:

$$\mathsf{M}_H\{z_1,\dots,z_N\} = \frac{1}{N-m}\sum_{z_k\leqslant c} z_k + \frac{m}{N-m}c,$$

где m – количество аргументов $z_k > c$. Если положим $c = \bar{z}_{\rho_\alpha}$, то

$$\frac{\partial \mathsf{M}_{H,\rho_\alpha}}{\partial z_k} = \begin{cases} \frac{1}{N-m} + \frac{m}{N-m}\frac{\partial \mathsf{M}_{\rho_\alpha}}{\partial z_k}, & \text{если } z_k < \bar{z}_{\rho_\alpha} \\ \frac{m}{N-m}\frac{\partial \mathsf{M}_{\rho_\alpha}}{\partial z_k}, & \text{если } z_k \geqslant \bar{z}_{\rho_\alpha}. \end{cases}$$

Но при этом, $\frac{\partial \mathsf{M}}{\partial z_1} + \dots + \frac{\partial \mathsf{M}}{\partial z_N} = \frac{N}{N-m}$.

4. Принцип минимизации дифференцируемых средних, нечувствительных к выбросам

Таким образом, в условиях выбросов, предлагается искать $\mathbf{c}^*$ и $\mathbf{S}^*$, минимизируя функционал

$$\mathcal{Q}(\mathbf{c}_1,...,\mathbf{c}_K;\mathbf{S}_1,...,\mathbf{S}_K) = \mathsf{M}\big\{D(\mathbf{x}_1;\mathbf{c}_1,...,\mathbf{c}_K;\mathbf{S}_1,...,\mathbf{S}_K),\dots,D(\mathbf{x}_N;\mathbf{c}_1,...,\mathbf{c}_K;\mathbf{S}_1,...,\mathbf{S}_K)\big\}.$$

В силу дифференцируемости $\mathsf{M}\{z_1,\dots,z_N\}$ искомые центры $\mathbf{c}_1^*,\dots,\mathbf{c}_K^*$ и матрицы $\mathbf{S}_1^*,\dots,\mathbf{S}_K^*$ являются решением системы нелинейных уравнений:

$$\begin{cases} z_k = D(\mathbf{x}_k;\mathbf{c}_1,\dots,\mathbf{c}_K;\mathbf{S}_1,\dots,\mathbf{S}_K), & k=1,\dots,N \\ \mathbf{v} = \nabla\mathsf{M}\{z_1,\dots,z_N\} & \\ \mathbf{c}_j = \frac{1}{V_j}\sum\limits_{k\in\mathbf{I}_j} v_k\mathbf{x}_k, & j=1,\dots,K \\ \mathbf{S}_j = \frac{1}{V_j}\sum\limits_{k\in\mathbf{I}_j} v_k(\mathbf{x}_k-\mathbf{c}_j)'(\mathbf{x}_k-\mathbf{c}_j), & j=1,\dots,K \end{cases} \tag{8}$$

Вектор весов $\mathbf{v}$ при $\mathbf{c}_j = \mathbf{c}_j^*$ и $\mathbf{S}_j = \mathbf{S}_j^*$ также можно использовать в качестве оценки значимости точек. Так как $v_1 + \dots + v_N = 1$, то выбросам будут соответствовать точки с наименьшими значениями весов.

Устойчивость по отношению к выбросам достигается за счет того, что веса точек, соответствующих выбросам, оказывается существенно меньше, чем веса точек, не являющихся выбросами. Важно также то, величина веса точки убывает по мере роста модуля разности между $\bar{z} = \nabla\mathsf{M}\{z_1,\dots,z_N\}$ и z_k. Такие свойства являются естественным следствием робастности оценок среднего значения.

Например, для $\mathsf{M}_\rho\{z_1,\dots,z_N\}$

$$v_k = \frac{\rho''(z_k-\bar{z}_\rho)}{\rho''(z_1-\bar{z}_\rho)+\dots+\rho''(z_N-\bar{z}_\rho)}.$$

При $\rho(r) = \sqrt{\varepsilon^2+r^2}-\varepsilon$

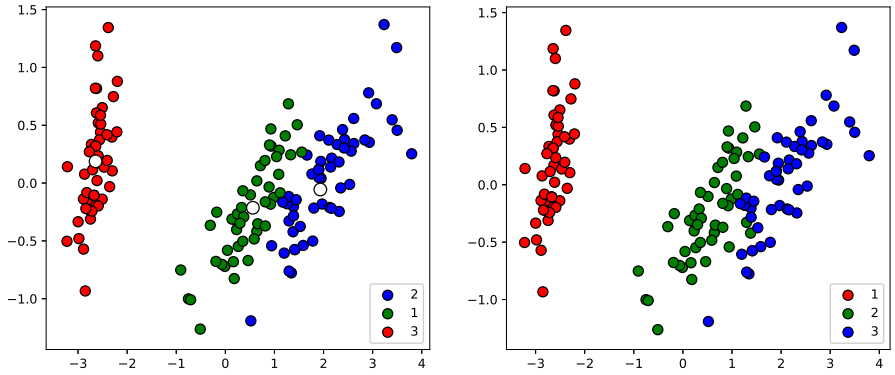
$$v_k < \left(\frac{\varepsilon^2+(z_k-\bar{z}_\rho)^2}{\varepsilon^2+(z_{(1)}-\bar{z}_\rho)^2}\right)^{3/2},$$

где $z_{(1)}$ – ближайшее значение к $\bar{z}_\rho$. Так что, v_k, будучи заключено между 0 и 1, быстро убывает по мере его удаления от $\bar{z}_\rho$.

Для $\mathsf{WM}_\rho\{z_1,\dots,z_N\}$, соответственно, получаем

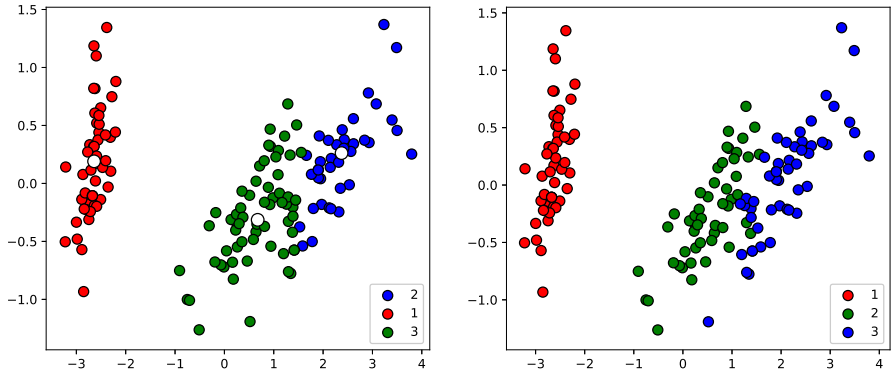
$$\begin{cases} \frac{1}{N} \leqslant v_k < \frac{1}{N} + \frac{m}{N}\left(\frac{\varepsilon^2+(z_k-\bar{z}_\rho)^2}{\varepsilon^2+(z_{(1)}-\bar{z}_\rho)^2}\right)^{3/2}, & \text{если } z_k < \bar{z}_\rho \\ v_k < \frac{m}{N}\left(\frac{\varepsilon^2+(z_k-\bar{z}_\rho)^2}{\varepsilon^2+(z_{(1)}-\bar{z}_\rho)^2}\right)^{3/2}, & \text{если } z_k \geqslant \bar{z}_\rho. \end{cases}$$

Здесь при $z_k \geqslant \bar{z}_\rho$ вес v_k быстро падает по мере удаления от $\bar{z}_\rho$. Веса $v_k > 1/N$ при $z_k < \bar{z}_\rho$ и быстро приближаются к $1/N$ по мере удаления значения z_k от $\bar{z}_\rho$.



а) б)

Рисунок 1. IRIS: Робастный алгоритм.



а) б)

Рисунок 2. IRIS: Классический алгоритм.

5. Алгоритм

Для поиска $\mathbf{c}_1^*, \ldots, \mathbf{c}_K^*$ и $\mathbf{S}_1^*, \ldots, \mathbf{S}_K^*$ применим итерационную схему, которая соответствует аналогу методу Якоби для решения системы нелинейных уравнений (8).

Начальные положения центров выбираются некоторым образом, например:

$$\begin{cases} \mathbf{c}_{j,0} = \dfrac{1}{N}\displaystyle\sum_{k=1}^{N} \mathbf{x}_k \\ \mathbf{S}_{j,0} = \mathbf{E}^{n\times n}, \end{cases}$$

где $\mathbf{E}^{n\times n}$ – единичная матрица $n \times n$.

(i) На t-ом шаге последовательно решаются два уравнения:

(a) Для каждого $j = 1, \ldots, K$ сначала решается следующее векторное уравнение для нахождения $\mathbf{c}_{j,t+1}$:

$$\mathbf{c}_j = \sum_{k\in\mathbf{I}_j} \frac{\partial \mathsf{M}\{z_1, \ldots, z_N\}}{\partial z_k}\mathbf{x}_k, \tag{9}$$

где $z_k = D(\mathbf{x}_k; \mathbf{c}_1, \ldots, \mathbf{c}_K; \mathbf{S}_{1,t}, \ldots, \mathbf{S}_{K,t})$.

(b) Затем для каждого $j = 1, \ldots, K$ решается следующее векторное уравнение для нахождения $\mathbf{S}_{j,t+1}$:

$$\mathbf{S}_j = \sum_{k \in \mathbf{I}_j} \frac{\partial \mathsf{M}\{z_1, \ldots, z_N\}}{\partial z_k} (\mathbf{x}_k - \mathbf{c}_{j,t+1})'(\mathbf{x}_k - \mathbf{c}_{j,t+1}), \tag{10}$$

где $z_k = D(\mathbf{x}_k; \mathbf{c}_{t+1}, \mathbf{S})$.

(ii) Шаг 1 повторяется до тех пор, пока $t < T$ (максимальное число итераций) или последовательность $\{\mathcal{Q}(\mathbf{c}_{t,1}, \ldots, \mathbf{c}_{t,K}; \mathbf{S}_{t,1}, \ldots, \mathbf{S}_{t,K})\}$ не сконцентрируется вокруг своей точки сгущения.

Множества индексов точек $\mathbf{I}_1, \ldots, \mathbf{I}_K$, соответствующие разбиению на кластеры, находятся перед решением систем уравнений. Дополнительное условие $|\mathbf{S}| = 1$ обычно добавляется, чтобы предотвратить вырождение корреляционной матрицы. Показатель масштаба $\sigma = |\mathbf{S}|$ можно затем оценить при помощи S-эстиматора [6].

Первое уравнение в системе имеет вид:

$$\mathbf{c} = F(\mathbf{c}).$$

Для его решения можно использовать итерационную процедуру:

$$\mathbf{c}_{t+1} = (1 - h)\mathbf{c}_t + hF(\mathbf{c}_t),$$

где $0 \leqslant h \leqslant 1$. Второе уравнение имеет аналогичный вид:

$$\mathbf{S} = G(\mathbf{S}).$$

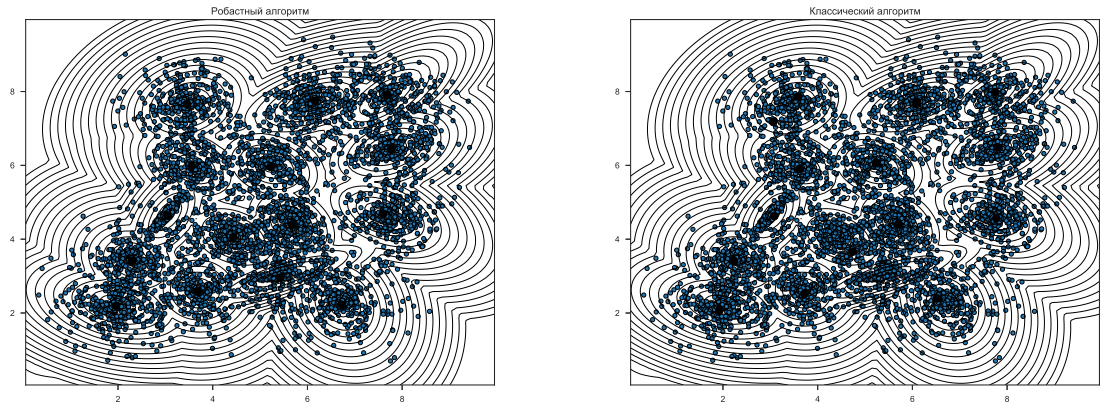
Для его решения можно использовать аналогичную итерационную процедуру:

$$\mathbf{S}_{t+1} = (1 - h)\mathbf{S}_t + hG(\mathbf{S}_t),$$

6. Примеры

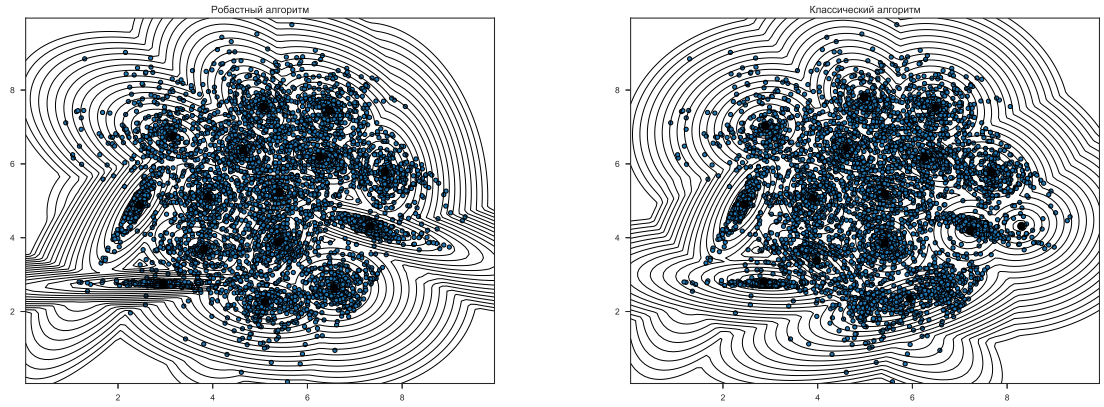
1. Расмотрим относительно простой и классический набор данных iris. Как правило, он используется для задач классификации. Мы же здесь попытаемся идентифицировать классы при помощи кластеризации, используя расстояния Махаланобиса вместо Евклидового. На Рис. 1 слева на а) представлен результат робастной кластеризации, справа на б) исходное разбиение на классы. Нетрудно убедиться, что можно получить разбиение, которое отличается от заданного в 3-х точках из 150-ти. Для сравнение на Рис. 2 приведено разбиение, полученное при помощи классического метода. Здесь отличие в 5-ти точках. Хотя преимущество минимальное, но с учетом того, что наилучшие надежные алгоритмы классификации по данным набора iris как раз дают 98% точности и выше, то это можно считать хорошим результатом. Он показывает, что применение робастного подхода к кластеризации на основе реалистичного набора признаков можно получать разбиения, которые практически соответствуют заданной классификации.

2. Рассмотрим наборы данных S3-S4 из [7, 8]. Они содержит 5000 точек, 15 кластеров. Среди наборов данных S1–S4 [8], в S3-S4 доля выбросов – *наибольшая*. Именно поэтому этот набор данных представляет наибольший интерес. Для сравнения приведем результаты применения стандартный метода k-means, но с учетом корреляционной матрицы. На Рис. 3 и 4 представлены результаты кластеризации для наборов S3 и S4, соответственно. На а) представлен результат работы робастного алгоритма, а на б) – классического.



а) б)

Рисунок 3. S3: Результаты робастного и классического алгоритмов.



а) б)

Рисунок 4. S4: Результаты робастного и классического алгоритмов.

7. Заключение

В настоящей работе рассмотривался вариант k-means, в котором вместо евклидова расстояния использовалось расстояние Махаланобиса. Предложенный новый подход к построению робастного варианта k-means на основе расстояния Махаланобиса, основанный на минимизации робастных дифференцируемых оценок среднего показал свою принципиальную устойчивость к искажениям в данных, по сравнению с классическим алгоритмом k-means. Связано это с тем, что применяемые в работе робастные оценки среднего ограничивают влияние на поиск положения центров кластеров точек, которые расположены на относительно больших расстояниях от них.

8. Благодарности

Работа выполнена при поддержке гранта РФФИ № 18–01–00050.

9. Литература

[1] Maronna, R.A. Robust M-Estimators of Multivariate Location and Scatter // Annals of Statistics. – 1976. – Vol. 4. – P. 51-67.
[2] Rousseeuw, P. High-breakdown estimators of multivariate location and scatter / P. Rousseeuw, M. Hubert // Robustness and Complex Data Structures, 2013. – P. 49-66.
[3] Maronna, R.A. Robust and efficient estimation of multivariate scatter and location / R.A. Maronna, V.J. Yohai // arxiv: 1504.03389, 2015.
[4] Шибзухов, З.М. О принципе минимизации эмпирического риска на основе усредняющих агрегирующих функций // Доклады РАН. – 2017 – Т. 476, № 5. – С. 495-499.
[5] Shibzukov, Z.M. Clustering based on the principle of finding centers and robust averaging functions of aggregation / Z.M. Shibzukov, M.A. Kazakov // Proceedings of V International Conference Information Technology and Nanotechnology, 2019. - P. 2014-2018.
[6] Davies, P.L. Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices // Annals of Statistics. – 1987. – Vol. 15. – P.1269-1292.
[7] Franti, P. K-means properties on six clustering benchmark datasets / P. Franti, S. Sieranoja // Applied Intelligence. – 2018. – 48(12). – P. 4743-4759.
[8] Clustering basic benchmark [Electronic Resourse]. – Access mode: http://cs.joensuu.fi/sipu/datasets/.

Robust k-means method based on minimizing differentiable estimates of mean insensitive to outliers

Z.M. Shibzukhov[1,2], M.A. Kazakov[1], D.P. Dimitrichenko[1]

[1]Institute of Mathematics and Informatics MPSU, Krasnoprudnaya str. 14, Moscow, Russia, 119991

[2]Institute of Applied Mathematics and Automation, KBNC RAS, Balkarova str. 2, Nalchik, Russia, 360002

Abstract. A new approach toconstructing avariant of the k-means clustering algorithm isproposed, in which the Mahalanobis distance is used instead of the Euclidean distance. It is based on minimizing di˙erentiable estimates of average values that are insensitive tooutliers. The examples show the possibility of stability of the proposed algorithm with respect tooutliers in the data.