Кластеризация на основе принципа поиска центров и робастных усредняющих агрегирующих функций

З.М. Шибзухов[1]

[1]Институт математики и информатики Московского педагогического государственного университета, Краснопрудная 4, Москва, Россия, 119991

Аннотация. Предлагается новый подход к робастной кластеризации на основе поиска центров кластеров. Он основан на минимизации робастных оценок средних и сумм функций псевдорасстояний до центров кластеров. Предложен алгоритм типа итеративного перевзвешивания для поиска центров кластеров. Приводятся примеры, показывающие устойчивость метода по отношению к большому количеству выбросов.

1. Введение

В основе одного классического подхода для разбиения на кластеры лежит процедура поиска центров кластеров. Центр кластера – это точка, от которой сумма расстояний до всех его точек минимальна. Разбиение на кластеры осуществляется по простому правилу: точка относится к кластеру, до центра которого расстояние минимально.

Мы предлагаем расширить принцип разбиения на кластеры. Для этого в определении центра кластера и для отнесения точки к кластеру будем использовать непрерывно-дифференцируемые усредняющие агрегирующие функции вместо обычной арифметической суммы и взятия минимума. Это позволяет расширить подход и предложить новые алгоритмы для поиска центров кластеров, предложенный в [1].

2. Алгоритм HCD

Пусть $\mathbf{S} \subset \mathbb{R}^n$ – открытое подмножество, $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathbf{S}$ – конечное множество, которое требуется разбить на K кластеров.

Пусть $d: \mathbf{S} \times \mathbf{S} \to \mathbb{R}_+$ – функция обобщенного расстояния между точками из $\mathbf{S}$, которая, по определению, удовлетворяет следующим требованиям:

1. $d_{\mathbf{c}}(\mathbf{x}) = d(\mathbf{x}, \mathbf{c})$ – строго выпуклая и дважды дифференцируемая на $\mathbf{S}$;
2. для любого $\mathbf{c} \in \mathbf{S}$: $\lim\limits_{\|\mathbf{x}\| \to \infty} d_{\mathbf{c}}(\mathbf{x}) = \infty$.

Определим классический метод разбиения на кластеры. Пусть $\mathbf{C}_1, \dots, \mathbf{C}_K$ — некоторое разбиение множества $\mathbf{X}$ на K кластеров, $y_j(\mathbf{x})$ — характеристическая функция -го кластера, $1 \le j \le K$:

$$y_j(\mathbf{x}) = \begin{cases} 1, & \text{если } \mathbf{x} \in \mathbf{C}_j \\ 0, & \text{иначе.} \end{cases}$$

Детерминированные характеристические функции кластеров, которые полностью задаются своими центрами, обычно определяются следующим образом:

$$y_j(\mathbf{x}) = \begin{cases} 1, & \text{если } d(\mathbf{x}, \mathbf{c}_j) = D(\mathbf{x}, \mathbf{c}_1, \dots, \mathbf{c}_K) \\ 0, & \text{иначе,} \end{cases}$$

где

$$D(\mathbf{x}, \mathbf{c}_1, \dots, \mathbf{c}_K) = \min\{d(\mathbf{x}, \mathbf{c}_1), \dots, d(\mathbf{x}, \mathbf{c}_K)\}.$$

Т.е. точка относится к тем кластерам, расстояние до которых минимально.

Задачу поиска кластеров, следуя [1], можно сформулировать как задачу оптимизации:

$$\mathbf{c}_1^*, \dots, \mathbf{c}_K^* = \arg\min_{\mathbf{c}_1,\dots,\mathbf{c}_K} Q(\mathbf{c}_1, \dots, \mathbf{c}_K), \quad (1)$$

где

$$Q(\mathbf{c}_1, \dots, \mathbf{c}_K) = \sum_{k=1}^{N} \nu_k D(\mathbf{x}_k, \mathbf{c}_1, \dots, \mathbf{c}_K), \quad (2)$$

величина $\nu_k \geq 0$ отражает значимость -ой точки, $\nu_1 + \dots + \nu_N = 1$.

Заметим, что в данной постановке используется взвешнное среднее вместо вместо обычной суммы или взвешенной суммы. Это не является помехой, когда количество точек N и веса точек являются заданными числовыми параметрами, т.к. любая взвешенная сумма легко заменяется на эквивалентное ей взвешенное среднее.

Так как

$$\nabla_{\mathbf{c}_j} Q = \sum_{k \in \mathbf{I}_j} \nu_k \nabla_{\mathbf{c}_j} d_k(\mathbf{c}),$$

где $d_k(\mathbf{c}_j) = d(\mathbf{x}_k, \mathbf{c}_j)$, $\mathbf{I}_j = \{k: \mathbf{x}_k \in \mathbf{C}_j\}$, то центр -ого кластера является решением следующей задачи оптимизации:

$$\mathbf{c}_j = \operatorname{argmin}_{\mathbf{c}} \sum_{k \in \mathbf{I}_j} \nu_k \, d_k(\mathbf{c}).$$

Классический алгоритм кластеризации на основе поиска центров кластеров HCD (*Hard Clustering with Distance-like functions*) представляет собой итерационный процесс уточнения положения центров $\mathbf{c}_1, \dots, \mathbf{c}_K$. Он представляет собой разновидность алгоритмов спуска по альтернативным направлениям, которые соответствуют центрам кластеров. Псевдокод алгоритма можно записать в следующей форме:

procedure HCD($\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\{\mathbf{c}_1^0, \dots, \mathbf{c}_K^0\}$)
  $\mathbf{c}_1, \dots, \mathbf{c}_K \leftarrow \mathbf{c}_1^0, \dots, \mathbf{c}_K^0$
  repeat
    for all $j = 1, \dots, K$ **do**
      $\mathbf{I}_j = \{k: y_j(\mathbf{x}_k) = 1\}$
    $\mathbf{c}_j \leftarrow \operatorname{argmin}_{\mathbf{c}} \sum_{k \in \mathbf{I}_j} \nu_k \, d_k(\mathbf{c})$,
    end for
  until положения центров не стабилизируются
  return $\mathbf{c}_1, \dots, \mathbf{c}_K$
end

Обычно $\nu_k = 1/N$. Нестандартные наборы весов могут возникать косвенным образом в рамках процедуры робастной кластеризации, когда $d(\mathbf{x}, \mathbf{c}) = \varrho(\| \mathbf{x} - \mathbf{c} \|^2)$. Робастная кластеризация может получаться при определенных условиях, когда $\varrho(r)$ растет существенно медленнее, чем линейная функция.

Например, когда $\varrho(r) = \sqrt{r}$, $\varrho(r) = (1 + r)^\gamma - 1$ ($\gamma < 1/2$) или даже $\varrho(r) = \ln(1 + r)$.

3. M-средние

В настоящей статье рассмотрим обобщения этой постановки задачи, исходя из факта, что min и среднее арифметическое являются примерами усредняющих агрегирующих функций [2,3]. Значительный класс таких функций – это -средние. Будем строить эти обобщения по следующей схеме [4].

Пусть $\rho(r)$ – выпуклая функция. Определим M-среднее, как решение следующей задачи:

$$\mathsf{M}_\rho\{r_1, \dots, r_m\} = \operatorname{argmin}_s \sum_{j=1}^{m} \rho\,(r_j - s).$$

Если $\rho(r)$ – строго выпуклая функция, то M_ρ – усредняющая агрегирующая функция [5,6]. Большинство известных функций для вычисления эмпирического среднего можно представить как -среднее. Если существует $\rho''(r)$, то

$$\frac{\partial \mathsf{M}_\rho}{\partial r_j} = \frac{\rho''(r_j - \overline{r})}{\rho''(r_1 - \overline{r}) + \cdots + \rho''(r_m - \overline{r})},$$

где $\overline{r} = \mathsf{M}_\rho\{r_1, \dots, r_m\}$. При этом, $\frac{\partial \mathsf{M}_\rho}{\partial r_1} + \cdots + \frac{\partial \mathsf{M}_\rho}{\partial r_m} = 1$.

4. Алгоритм IR-HCD-M1

Рассмотрим *первое* обобщение путем замены взвешенного среднего арифметического на M_ρ:

$$Q(\mathbf{c}_1, \dots, \mathbf{c}_K) = \mathsf{M}_\rho\{D(\mathbf{x}_1), \dots, D(\mathbf{x}_N)\}.$$

Такая замена имеет смысл, когда среди точек есть выбросы, которые могут существенно сместить значение функционала, если их наберется достаточное большое количество. Из-за такого смещения как раз и проявляется смещение центров искомых кластеров.

Градиент Q по $\mathbf{c}_j$ $(1 \le j \le K)$ имеет вид:

$$\nabla_{\mathbf{c}_j} Q = \sum_{k \in \mathbf{I}_j} \frac{\partial \mathsf{M}_\rho\{D(\mathbf{x}_1), \dots, D(\mathbf{x}_N)\}}{\partial r_k} \nabla_{\mathbf{c_j}} d_k(\mathbf{c}_j),$$

где $\mathbf{I}_j = \{k : y_{kj} = 1\}$.

Для решения задачи поиска центров, минимизирующих целевой функционал можно построить процедуру спуска по альтернативным направлениям – градиентам $\nabla_{\mathbf{c}_j} Q$. Она также представляет собой разновидность процедуры итеративного перевзвешивания, отличаясь от алгоритма IR-KMeans способом пересчета весов. Псевдокод ее можно засписать следующим образом:

procedure IR-HCD-M1($\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\{\mathbf{c}_1^0, \dots, \mathbf{c}_K^0\}$)
  $\mathbf{c}_1, \dots, \mathbf{c}_K \leftarrow \mathbf{c}_1^0, \dots, \mathbf{c}_K^0$
  repeat
    for all $j = 1, \dots, K$ **do**
      for all $k \in \mathbf{I}_j$ **do** $v_k = \frac{\partial \mathsf{M}_\rho\{D(\mathbf{x}_1), \dots, D(\mathbf{x}_N)\}}{\partial r_k}$
      $\mathbf{c}_j \leftarrow \operatorname{argmin}_{\mathbf{c}} \sum_{k \in \mathbf{I}_j} v_k \, d_k(\mathbf{c})$
    end for
  until положения центров не стабилизируются
  return $\mathbf{c}_1, \dots, \mathbf{c}_K$
end

По окончании работы алгоритма получаются веса точек $v_1^*, \dots, v_N^*$. Найденные центры при определенных условиях также являются решением задачи (1)–(2) с весами $v_1^*, \dots, v_N^*$.

Если $d(\mathbf{x}, \mathbf{c}) = \| \mathbf{x} - \mathbf{c} \|^2$, то алгоритм IR-HCM-M1 превращается в аналог IR-KMeans, в котором веса $v_{kj} = \frac{\partial \mathsf{M}_\rho\{d_{k1}, \dots, d_{kK}\}}{\partial r_j}$, где $d_{kj} = \| \mathbf{x}_k - \mathbf{c}_j \|^2$.

Если выбрать ρ так, чтобы M_ρ определяла робастную функцию -среднего, то можно построить робастный алгоритм кластеризации.

4.1 Иллюстративные примеры

Для иллюстрациии возможностей алгоритма IR-KMeans-M1 рассмотрим ряд примеров. Они наглядно демонстрируют его способность находить центры, которые лежат достаточно близко к настоящим центрам кластеров в условиях, когда данные содержат выбросы или когда ищутся центры не всех кластеров, а лишь некоторая часть из них. Если в первом случае причиной смещения найденных центров являются выбросы, то во втором случае причиной смещения являются избыточные кластеры.

Пример 1. В этом примере сравнивается применение алгоритмов KMeans и IR-KMeans-M1 с -средним $\mathsf{M}_{\rho_{\alpha,\varepsilon}}$, где $\rho_\varepsilon(r) = \sqrt{\varepsilon^2 + r^2}$, $\varepsilon = 0.001$, для поиска центров двух кластеров. Точки искусственно сгенерированного набора данных принадлежат двум кластерам с нормальным распределением расстояний точек кластера от своего центра. Дополнительно добавлено облако "выбросов", которые расположено на удалении от центров первых двух кластеров и имеет большую дисперсию в распределении расстоянии точек от его центра.
В первом случае, первые два кластера содержат 200 точек, облако "выбросов" содержит 100 точек (33%). Обычный метод KMeans выдает центры кластеров, смещенные в сторону выбросов. Алгоритм IR-KMeans-M1 с $\alpha = 0.4$. Во втором случае в облаке "выбросов" содержится уже 200 точек (50%), также несколько увеличен разброс его точек. Для робастной кластеризации применен алгоритм IR-KMeans-M1 с $\alpha = 0.3$. Найденные центры также лежат внутри своих кластеров, что позволяет отсечь выбросы при помощи распределения значений $\{D(\mathbf{x}_k, \mathbf{c}_1^*, \mathbf{c}_2^*): k = 1, \dots, N\}$. Результаты представлены на Рисунке 1.

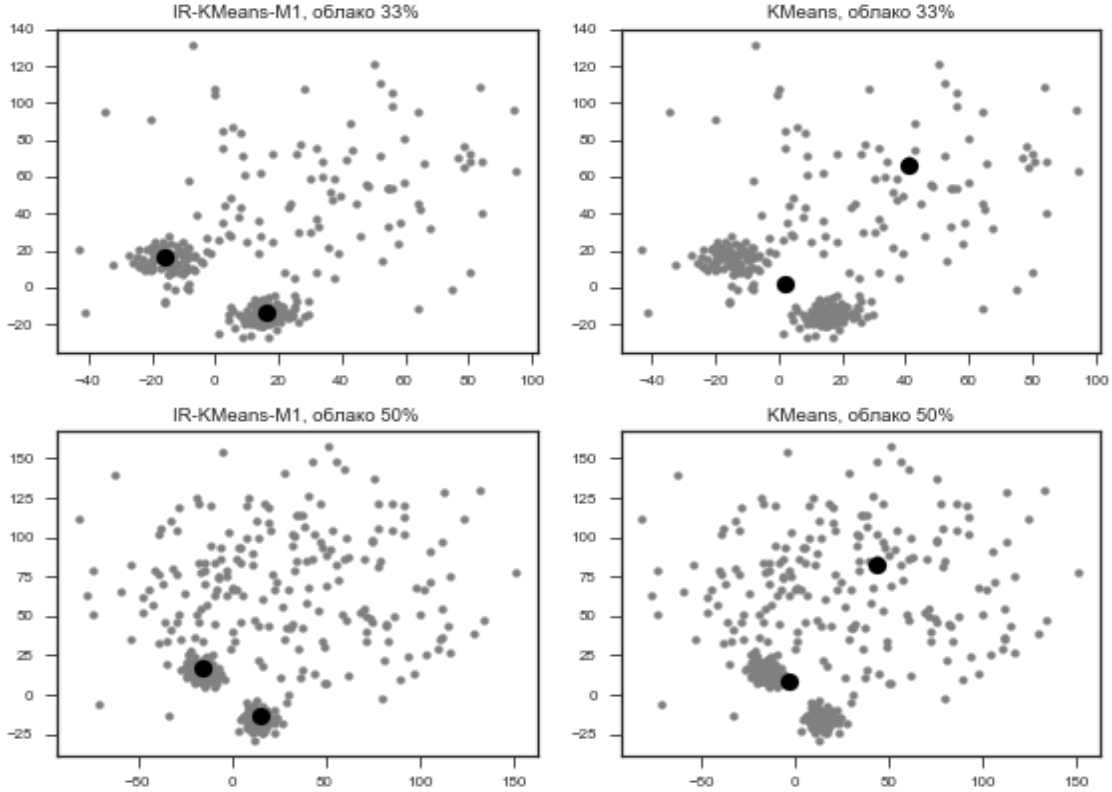

Рисунок 1. Применение алгоритмов KMeans и IR-KMeans-M1 в примере 1.

Пример 2. В этом примере случайно сгенерированы 8 кластеров, так что в каждом кластере точки распределены по нормальному закону, кластеры не пересекаются между собой. Алгоритм IR-KMeans-M1 применяется для кластеризации для $K = 2,3,4,5,6,7$ кластеров, соответственно. При этом, использовалось использовались M-средние $\mathsf{M}_{\rho_{\alpha,\varepsilon}}$,где $\rho_\varepsilon(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon$, $\varepsilon = 0.001$ с $\alpha = 0.15,0.20,0.25,0.30,0.35,0.40$, соответственно. Во всех случаях найденные центры попадают внутрь кластеров. Результаты представлены на Рисунке 2.

5. Алгоритм IR-HCD-MM

Заметим, что усредняющую функцию min можно апроксимировать при помощи -средних. Например, на основе использовния дифференцируемого приближения -квантиля при достаточно малых α. Его можно построить при помощи функции $\mathsf{M}_{\rho_{\alpha,\varepsilon}}$, где

$$\rho_{\alpha,\varepsilon}(r) = \begin{cases} \alpha\rho_\varepsilon(r), & \text{если } r > 0 \\ \dfrac{\alpha}{2}\rho_\varepsilon(0_+) + \dfrac{1-\alpha}{2}\rho_\varepsilon(0_-), & \text{если } r = 0 \\ (1-\alpha)\rho_\varepsilon(r), & \text{если } r < 0, \end{cases}$$

$\rho_\varepsilon(r)$ – такая функция, что

1) $\lim\limits_{\varepsilon\to 0}\rho_\varepsilon(r) = |r|$;
2) $\lim\limits_{\varepsilon\to 0}\rho'_\varepsilon(r) = \operatorname{sign} r$;
3) $\lim\limits_{\varepsilon\to 0}\rho''_\varepsilon(r) = \delta(r)$ – δ-функция Дирака.

Например:

1) $\rho_\varepsilon(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon$;
2) $\rho_\varepsilon(r) = |r| - \varepsilon\ln(\varepsilon + |r|) + \varepsilon\ln\varepsilon$.

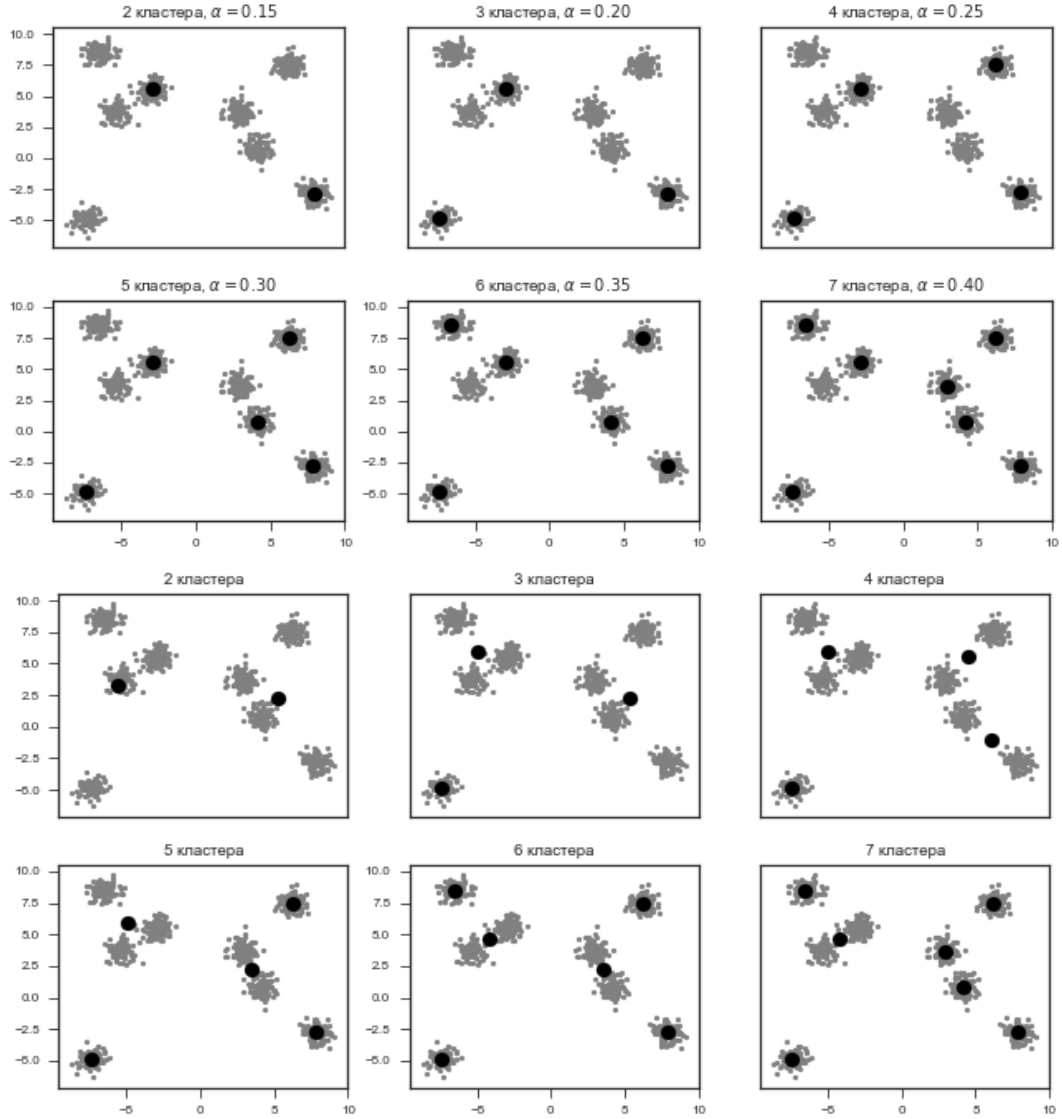

Рисунок 2. Применение алгоритма IR-KMeans-M1 и KMeans в примере 2.

Такие приближения предпочтительны в тех случаях, когда точку нельзя уверенно отнести к центру одного кластера из-за того, то еще один или несколько центров других кластеров расположены очень близко.

Используя апроксимации минимума при помощи M_χ для некоторой функции $\chi = \rho_{\alpha,\varepsilon}$ построим *второе* обобщение (2), когда

$$D(\mathbf{x}) = \mathsf{M}_\chi\{d(\mathbf{x},\mathbf{c}_1),\dots,d(\mathbf{x},\mathbf{c}_K)\}.$$

Объединим оба способа, изложенные выше. Теперь задача поиска центров имеет вид:

$$\mathbf{c}_1^*,\dots,\mathbf{c}_K^* = \arg\min_{\mathbf{c}_1,\dots,\mathbf{c}_K} Q(\mathbf{c}_1,\dots,\mathbf{c}_K),$$

где

$$Q(\mathbf{c}_1,\dots,\mathbf{c}_K) = \mathsf{M}_\rho\{D(\mathbf{x}_1),\dots,D(\mathbf{x}_N),$$

Градиент Q по $\mathbf{c}_j$ $(1 \le j \le K)$ имеет вид:

$$\operatorname{grad}_{\mathbf{c}_j} Q = \sum_{k=1}^{N} \frac{\partial \mathsf{M}_\rho\{d_{1j},\dots,d_{Nj}\}}{\partial r_k} \frac{\partial \mathsf{M}_\chi\{d_{k1},\dots,d_{kK}\}}{\partial r_j} \operatorname{grad}_{\mathbf{c_j}} d(\mathbf{x}_k,\mathbf{c}_j),$$

где $d_{kj} = d(\mathbf{x}_k,\mathbf{c}_j)$.

Псевдокод процедуры итеративного первзвешивания для поиска центров $\mathbf{c}_1^*,\dots,\mathbf{c}_K^*$ можно записать следующим образом:

procedure IR-HCD-MM($\{\mathbf{x}_1,\dots,\mathbf{x}_N\}$, $\{\mathbf{c}_1^0,\dots,\mathbf{c}_K^0\}$)
  $\mathbf{c}_1,\dots,\mathbf{c}_K \leftarrow \mathbf{c}_1^0,\dots,\mathbf{c}_K^0$
  repeat
    for all $k = 1,\dots,N$ **do**
      $(v_{k1},\dots,v_{kK}) = \operatorname{grad}\mathsf{M}_\chi\{d_k(\mathbf{c}_1),\dots,d_k(\mathbf{c}_K)\}$
    for all $j = 1,\dots,K$ **do**
      $(v_{1j},\dots,v_{Nj}) = \operatorname{grad}\mathsf{M}_\rho\{d_1(\mathbf{c}_j),\dots,d_N(\mathbf{c}_j)\}$
    for all $j = 1,\dots,K$ **do**
      $\mathbf{c}_j \leftarrow \operatorname{argmin}_{\mathbf{c}} \sum_{k=1}^{N} v_{kj}\, v_{kj} d_k(\mathbf{c})$
  until положения центров не стабилизируются
  return $\mathbf{c}_1,\dots,\mathbf{c}_K$
end

Заметим, что $v_{k1} + \dots + v_{kK} = 1$. Это прозволяет интерпретировать величины $v_{k1},\dots,v_{kK}$ как степень принадлежности точки $\mathbf{x}_k$ к классам с метками $1,\dots,K$, соответственно, получая нечеткое разбиение на кластеры. При этом, нечеткая функция принадлежности к кластерам строится на основе выбранной дифференцируемой усредняющей функции M_χ следующим образом:

$$y_1(\mathbf{x}),\dots,y_K(\mathbf{x}) = \operatorname{grad}\mathsf{M}_\chi\{d(\mathbf{x},\mathbf{c}_1),\dots,d(\mathbf{x},\mathbf{c}_K)\}.$$

Алгоритм IR-HCM-MM представляет вариант обобщения алгоритма SKM [1]. Там же (в [1]) было показано, что SKM обобщает уже известные алгоритмы кластеризации, как такие FCM [7], EM [8], DA [9], Bergman Soft Clustering [10]. Соответственно, если вместо $\mathsf{M}\chi$ использовать средние по Колмогорову, то можно также построить алгоритм, которые непосредственно обобщает SKM.

6. Заключение

Таким образом, применение -средних позволило построить новые процедуры поиска центров кластеров, которые обобщают HCD и SKM, позволяя использовать широкий спектр методов поиска среднего значения как для четкого, так и для нечеткого отнесения точек к кластерам.

7. Литература

[1] Teboulle, M. A Unified Continuous Optimization Framework for Center-Based Clustering Method // Journal of Machine Learning Research. – 2007 – Vol. 8. – P. 65-102.

[2] Mesiar, R. Aggregation functions: A revision / R. Mesiar, M. Komornikova, A. Kolesarova, T. Calvo // Fuzzy Sets and Their Extensions: Representation, Aggregation and Models. – Springer, Berlin, Heidelberg, 2008.

[3] Grabich, M. Aggregation Functions / M. Grabich, J.-L. Marichal, E. Pap. – Series: Encyclopedia of Mathematics and its Applications. Cambridge University Press. – 2009. – Vol. 127.
[4] Шибзухов, З.М. О принципе минимизации эмпирического риска на основе усредняющих агрегирующих функций // Доклады РАН. – 2017. – Т. 476, № 5. – С. 495-499.
[5] Calvo, T. Aggregation functions based on penalties / T. Calvo, G. Beliakov // Fuzzy Sets and Systems. – 2010. – Vol.161(10). – P. 1420-1436.
[6] Beliakov, G. A Practical Guide to Averaging Functions / G. Beliakov, H. Sola, T. Calvo. – Springer, 2016.
[7] Bezdek, J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. – Plenum Press, N. York, 1981.
[8] Duda, R.O. Pattern Classification / R.O. Duda, P.E. Hart, D.G. Stork. – John Wiley & Sons Inc., 2001.
[9] Rose, K. A deterministic annealing approach to clustering / K. Rose, E. Gurewitz, C.G. Fox // Pattern Recognition Letters. – 1990. – Vol. 11(9). – P. 589-594.
[10] Banerjee, A. Clustering with Bregman Divergences / A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh // Journal of Machine Learning Research. – 2005. – Vol. 6. – P. 1705-1749.

Благодарности
Работа выполнена при поддержке гранта РФФИ 18-01-00050.

Center based clustering with robust averaging aggregation functions

Z.M. Shibzukhov[1]

[1]Institute of mathematics and informatics, Krasnoprudnaya 4, Moscow, Russia, 119991

Abstract. A new approach to robust clustering based on the search for cluster centers is proposed. It is based on minimizing robust estimates of the averages and the sum of the functions of pseudo-distances to cluster centers. An algorithm of iterative reweighing type for finding cluster centers is proposed. Examples are given showing the stability of the method with respect to a large number of emissions.