УДК 681.3: 519.68

В.И. Загребнюк, В.Ю. Кумыш

МЕТОД ВЫБОРА НАЧАЛЬНЫХ ПРИБЛИЖЕНИЙ ЦЕНТРОВ КЛАСТЕРОВ ДЛЯ АЛГОРИТМА *K*-СРЕДНИХ

Разрабатывается метод поиска начальных приближений центров кластеров для алгоритма К-средних, который позволяет сократить количество итераций алгоритма К-средних в 2,9 раза по сравнению с существующими методами. Нулевые кластеры в ходе обработки данных отсутствуют. Вычислительная сложность процедуры кластеризации в целом у предложенного метода меньше, чем у существующих методов, так как алгоритм К-средних в процедуре выбора начальных приближений не используется. В отличие от других методов ошибка кластеризации контролируется на стадии выбора начальных приближений. Также точность кластеризации увеличивается за счет того, что начальные приближения выбираются по всем существенным признакам исходного набора данных и выполняется замена полученных значений начальных приближений ближайшими элементами исходного набора данных.

Введение

Кластерный анализ используется в области Data Mining, статистическом анализе данных, цифровой обработке сигналов для кластеризации и сегментации изображений в пространстве цветов, для квантования цветов с целью уменьшения цветовой избыточности и т. д.

Задача кластерного анализа заключается в разбиении множества X, которое содержит значения существенных признаков объектов или данных x_j, $j \in [1, m]$, на K подмножеств или кластеров $C_i, i \in [1, K]$, таких, что $X = \bigcup_{i=1}^{K} C_i$, причем $C_j \bigcap C_i = \emptyset, \forall i \neq j$. Множество X рассматривают как n-мерное пространство, в котором объекты или данные отображаются в виде точек $(x_{j1}, x_{j1}, \ldots, x_{jn})$ или векторов $\vec{x}_j = (x_{j1}, x_{j1}, \ldots, x_{jn})$, где x_{jk} – существенные признаки j-го объекта.

Для кластерного анализа наиболее часто используется алгоритм *K*-средних [1] и его модификации. Поскольку *K*-средних – это эвристический алгоритм, результаты его выполнения очень чувствительны к выбору начальных приближений центров кластеров [2, 3]. Одна из наиболее сложных и актуальных проблем практического использования алгоритма *K*-средних – проблема выбора начальных приближений. При неудачном выборе начальных приближений могут возникнуть пустые кластеры, а в случае квантования изображений еще и искажения цветов. Для практических задач статистического анализа данных, цифровой обработки сигналов и изображений это недопустимо. От выбора начальных значений центров кластеров существенно зависит также и вычислительная сложность алгоритма. Поэтому актуальна задача разработки методов выбора начальных приближений центров кластеров для алгоритма
K-средних.

Цель данной работы состоит в разработке такого метода выбора начальных приближений, который, с одной стороны, уменьшит количество итераций при выполнении алгоритма *K*-средних, а с другой – минимизирует и позволит контролировать ошибку кластеризации на стадии выбора начальных приближений.

1. Анализ существующих методов выбора начальных приближений центров кластеров для алгоритма *K*-средних

Известно, что задачи, которые решаются с использованием *K*-средних, относятся к задачам дискретной оптимизации и заключаются в поиске локальных экстремумов для K кластеров, которые обеспечивают минимизацию дисперсии σ^2, а именно

$$\min \sigma_i^2 = \min\left[\frac{1}{m_i}\cdot\sum_{j=1}^{m_i}\left(x_j^{(i)} - c_i\right)^2\right],$$

где c_i – центр кластера или центроид; $x_j^{(i)}$ – элемент множества X, который принадлежит кластеру C_i; m_i – мощность кластера C_i.

Выполнение алгоритма *K*-средних начинается с выбора заданного количества K начальных приближений центров кластеров или центроидов.

Наиболее простыми и общеизвестными методами выбора начальных приближений являются:

– случайный выбор начальных значений центроидов, в качестве начальных приближений произвольно выбирается K элементов из множества X, подлежащего кластеризации;

– выбор из X первых K элементов $x_i \in X, i = 1, ..., K$.

Среди известных методов выбора начальных приближений алгоритма *K*-средних можно выделить следующие:

Предварительная кластеризация с целью выбора начальных приближений. Множество X случайным образом делят на n подмножеств $\widetilde{X}_i, i \in [1, n]$, таких, что мощность их объединения равняется мощности множества X. Каждое $\widetilde{X}_i$ кластеризуют с помощью *K*-средних. Таким образом получают n вариантов начальных приближений. Из этих вариантов выбирают тот, который обеспечивает наименьшую ошибку кластеризации [3].

Рекурсивное нахождение начальных приближений с использованием алгоритма K-средних и метода K-деревьев. Для построения *K*-дерева применяется разделяющая гиперплоскость, перпендикулярная к оси с максимальной дисперсией. Ось с максимальной дисперсией определяется с помощью преобразования Карунена – Лоева. Построение *K*-дерева, разделение на кластеры осуществляется до тех пор, пока не получают заданное количество листов, равное заданному количеству кластеров. Центры листов используют в качестве начальных приближений [4].

Выбор начальных приближений на основании предварительного анализа каждого из существенных признаков $x_i, i \in [1, n]$, *отдельно* [5]. Для этого сначала рассчитываются средние значения $\hat{x}_i$ и среднеквадратическое отклонение σ_i каждого из существенных признаков. Множество значений каждого из существенных признаков разбивают на процентили $p_1, p_2, \ldots, p_K$ в соответствии с уровнями $\left(-\infty, (2j-1)/2K\right], j \in [1, K]$. В пределах каждого процентиля рассчитываются средние значения

$$\hat{x}_j^{(i)} = p_j \cdot \sigma_i + \hat{x}_i$$

для каждого из существенных признаков. Полученные $\hat{x}_j^{(i)}$ используются в качестве начальных приближений для кластеризации значений каждой компоненты с помощью алгоритма *K*-средних. После кластеризации определяют принадлежность данных полученным кластерам. При этом данное x_l принадлежит кластеру $C_{p_j}^i$, если значение его существенного признака принадлежит p_j. Для каждого p_j находят $C_{p_j} = \bigcup_i C_{p_j}^i$. Для каждого такого объединения вычисляют средние значения и используют их в качестве начальных приближений центроидов.

Выбор медиан [6]. В пределах множества данных X для каждого из существенных признаков вычисляется дисперсия. Для определения начальных приближений выбирается признак с максимальной дисперсией $x_{\sigma^2 \max}$. Множество значений $x_{\sigma^2 \max}$ сортируют, например, в порядке возрастания и делят на K полуинтервалов. В пределах каждого из этих полуинтервалов

находят медианы μ_j. Для каждой медианы в множестве X устанавливается соответствующее данное $x_{\mu j}$, у которого значение существенного признака $x^{(i)}_{\sigma^2 \max} = \mu_j$. Выбранные таким способом $x_{\mu j}$ используются в качестве начальных приближений для алгоритма *K*-средних.

Выбор начальных приближений, в котором рассматриваются только значения одного существенного признака $x_{\sigma^2 \max}$, имеющего наибольшую дисперсию σ^2 [7]. В общих чертах, выбор начальных приближений осуществляется с использованием итерационной процедуры, которая заключается в следующем. Данные сортируются в порядке возрастания значений существенного признака $x_{\sigma^2 \max}$. Вычисляются расстояния $d_i^2(x_i, x_{i+1}) = (x_i - x_{i+1})^2$, формируется множество $D = \{d_i^2 : i \in [1, n-1]\}$ и вычисляется сумма расстояний $D_{sum} = \sum_{i=1}^{n} d_i^2$. На первом этапе множество X делится на два кластера C_1, C_2 по оси $x_{\sigma^2 \max}$ точкой $D_{c0} = D_{sum}/n$. Вычисляется δ-ошибка кластеризации как разница ошибки кластеризации для X и суммы ошибок кластеризации для C_1, C_2. Дальнейшее разделение на кластеры осуществляется так, чтобы δ-ошибка уменьшалась, пока не будет получено заданное количество кластеров. Центроиды этих кластеров используются в качестве начальных приближений для алгоритма *K*-средних. Этот метод имеет наименьшую вычислительную сложность по сравнению с рассмотренными выше алгоритмами.

Сходимость алгоритма *K*-средних в общем случае пока что невозможно доказать [7, 8]. Поэтому для оценки качества кластеризации (ошибки кластеризации) используют, например, параметр SSE (Sum of Square Error) [6, 7]:

$$\mathrm{SSE} = \sum_j \sum_i \left(x_{ji} - c_j\right)^2,$$

где j – индекс кластера; i – индексы данных, которые принадлежат j-му кластеру.

Сравнивая SSE для разных вариантов кластеризации, выбирают тот, у которого SSE наименьшая.

В области цифровой обработки изображений наиболее часто для оценки ошибки кластеризации используются параметры MSE (Mean Squared Error) и PSNR (Peak Signal-to-noise Ratio):

$$\mathrm{MSE} = \frac{1}{MN} \sum_j^M \sum_i^N \left(x_{ji} - c_j\right)^2;$$

$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{2^r - 1}{\mathrm{MSE}},$$

где M – количество кластеров; N – количество пикселей в кластере; r – количество бит в пикселе.

Для определения того, насколько удачным является предложенный в [7] метод выбора начальных приближений, использовалось сравнение SSE для двух вариантов кластеризации: случайного выбора начальных приближений и выбора медиан. Предложенный в [7] метод имеет меньшую SSE, но в нем отсутствует сравнительная оценка сложности алгоритма *K*-средних, поэтому невозможно определиться с эффективностью его использования в случаях большого количества кластеров и особенно кластеризации изображений в пространстве цветов.

При применении методов выбора начальных приближений центров кластеров, описанных в [6, 7], в большинстве случаев ошибка кластеризации меньше, чем при использовании случайного выбора начальных приближений. Тем не менее встречаются такие наборы данных, для

которых использование случайного выбора начальных приближений дает меньшую ошибку кластеризации.

Среди недостатков указанных методов выбора начальных приближений можно назвать такие:

– использование алгоритма *K*-средних в процедурах выбора начальных приближений, что значительно увеличивает вычислительную сложность процедуры кластеризации в целом;

– одномерный выбор начальных приближений для многомерных данных, что в действительности может привести в лучшем случае не только к неудачному выбору начальных приближений, но и к увеличению ошибки кластеризации. Например, данные множества X, которые вообще принадлежат разным кластерам, но имеют одинаковые значения на оси $x_{\sigma\max}$, будут рассматриваться как одно данное. Это в итоге приведет к неправильной кластеризации, а именно к тому, что некоторые кластеры вообще не будут выявлены;

– наличие нулевых кластеров, что может привести к потере сходимости алгоритма.

2. Метод выбора начальных приближений в пространстве Карунена – Лоева

Разработанный метод выбора начальных приближений позволяет, с одной стороны, уменьшить количество итераций при выполнении алгоритма *K*-средних, а с другой – минимизировать и контролировать ошибку кластеризации на стадии начальных приближений.

Пусть пространство существенных признаков имеет n измерений. На первом этапе для множества X вычисляется ковариационная матрица Σ, элементы которой – ковариации – вычисляются по формуле

$$\sigma_{ij} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \mu_i)(x_j - \mu_j), \tag{1}$$

где μ_i – среднее значение i-го существенного признака.

Собственные значения находятся из решения характеристического уравнения

$$\Sigma - \lambda E = 0, \tag{2}$$

где E – единичная матрица.

Поскольку ковариационная матрица симметрична, решение этого характеристического уравнения находится с использованием метода Якоби, что дает возможность найти не только множество собственных значений $\lambda_i, i \in [1, n]$, но и множество собственных векторов $\vec{v}_i$. Известно, что симметричная матрица размерности $n \times n$ имеет n разных действительных значений и, соответственно, собственные векторы являются ортонормированными. Сформируем из собственных векторов матрицу

$$V = \left|v_{ij}\right| = \begin{vmatrix} v_{11} & v_{12} & \dots & v_{1n} \\ v_{21} & v_{22} & \dots & v_{2n} \\ \dots & \dots & \dots & \dots \\ v_{n1} & v_{n2} & \dots & v_{nn} \end{vmatrix}, \tag{3}$$

где i – индекс компоненты собственного вектора; j – номер собственного вектора.

Используя матрицу собственных векторов, перейдем в новую систему координат пространства Корунена – Лоева:

$$\tilde{x} = V^T x \ . \tag{4}$$

Найдем нормированные значения собственных чисел:

$$\tilde{\lambda}_i = \frac{\lambda_i}{\sqrt{\sum_{j=1}^{n} \lambda_j^2}}. \quad (5)$$

Используя нормированные собственные числа, выполним преобразование сжатия:

$$y = \Lambda \tilde{x}. \quad (6)$$

В результате данные из множества X группируются в пределах вытянутого гиперэллипсоида:

$$\frac{\left(y_{j1} - \hat{y}_1\right)^2}{\lambda_1} + \frac{\left(y_{j2} - \hat{y}_2\right)^2}{\lambda_2} + \ldots + \frac{\left(y_{jn} - \hat{y}_n\right)^2}{\lambda_n} \le R^2,$$

наибольшая полуось которого равна $R\sqrt{\lambda_m}$, где $\lambda_m = \max_i \{\lambda_i\}$. В результате такого преобразования получаем множество декоррелированных данных, что позитивно влияет на качество кластеризации.

Найдем два вектора $y_{\min}$ и $y_{\max}$, имеющих соответственно минимальное и максимальное значения компоненты y, которая соответствует максимальному собственному значению, тогда

$$\frac{y_{\max} - y_{\min}}{2} = 2R\sqrt{\lambda_m}.$$

Выполним преобразование смещения $\tilde{y} = y - y_{\min}$ и преобразование поворота так, чтобы вектор $\tilde{y}$ совпал с осью, которая соответствует максимальному собственному значению:

$$z = G\tilde{y},$$

где компоненты матрицы G – направляющие косинусы вектора $\tilde{y}_{\max}$.

На следующем этапе выполняется итерационный поиск начальных приближений центров кластеров. Обозначим ось, которая соответствует λ_m, через $0z_\sigma$, тогда все данные, которые подлежат кластеризации, будут находиться в промежутке $\left[0, z_\sigma^{\max}\right]$. В данном промежутке вычисляется среднее значения $\hat{z}_i$:

$$\hat{z}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} z_\sigma^j, \quad (8)$$

где i – номер итерации, $i = 1$; N_i – количество элементов в кластере, который делится (на первой итерации N_i соответствует общему количеству элементов в множестве X).

Полученным средним значением ось z_σ делится на два диапазона: $\left[0, \hat{z}_1\right)$ и $\left[\hat{z}_1, z_\sigma^{\max}\right]$, а множество X – на два кластера: C_1, C_2. Для полученных кластеров (диапазонов) вычисляется взвешенное количество элементов k_i' и длина D_i:

$$k_i' = \frac{k_i}{N}, \tag{9}$$

где k_i – количество элементов кластера C_i; N – общее количество элементов в множестве X.

Далее вычисляется среднее значение $\hat{z}_i$, $i = i + 1$, в кластере (диапазоне) с максимальным произведением $k_i' \cdot D_i$. Средним значением $\hat{z}_i$ данный кластер (диапазон) делится на два. Выбор следующего диапазона с максимальным произведением $k_i' \cdot D_i$ и его деление повторяются, пока не будет найдено K диапазонов. Для найденных диапазонов рассчитываются средние значения, которые и составляют начальные приближения. Следует подчеркнуть, что уже при этом максимальная ошибка кластеризации не будет превышать $\lambda_k < 1$, где $\lambda_k = \max_i\left(\{\lambda_i\} \setminus \lambda_m\right)$.

Для того чтобы улучшить выполнение алгоритма *K*-средних и увеличить точность кластеризации, в каждом диапазоне вычисляются средние значения для всех существенных признаков множества X и выполняется замена полученных значений начальных приближений ближайшими элементами из преобразованного множества X.

Полученные начальные приближения могут использоваться для выполнения алгоритма *K*-средних в преобразованном пространстве. Для кластеризации в исходном пространстве существенных признаков необходимо выполнить обратные преобразования поворота $\tilde{y}_i = G^{-1}\bar{z}_i$, смещения $y_i = \tilde{y}_i + y_{\min}$ и растяжения $\tilde{x}_i = \Lambda^{-1} y_i$.

Анализ результатов выбора начальных приближений для тестовых изображений со специально составленной базы [10] показал, что в среднем ошибка кластеризации на этапе выбора начальных приближений не превышает $\lambda_k \approx 0{,}009$.

3. Результаты апробации метода

Для того чтобы реализовать предложенный метод выбора начальных приближений центров кластеров, было разработано программное обеспечение для кластеризации цветных цифровых изображений с помощью алгоритма *K*-средних.

Было проведено сравнение результатов выполнения алгоритма *K*-средних для случайного выбора начальных приближений и с помощью разработанного метода по таким характеристикам, как наличие пустых кластеров и количество итераций. При случайном выборе начальных приближений во время обработки данных появились нулевые кластеры, причем их количество росло по мере увеличения общего количества кластеров (табл. 1).

При выборе начальных приближений с использованием предложенного метода нулевые кластеры отсутствуют.

Таблица 1

Количество пустых кластеров
при случайном выборе начальных приближений

Название файла	Количество нулевых кластеров		
	для 25 кластеров	для 50 кластеров	для 75 кластеров
3096	1	41	286
26031	0	3	10
35010	0	17	0
35091	0	0	39
41004	0	1	24
41025	1	3	74
54005	0	1	3
78004	1	1	3
86000	0	41	36
144067	1	1	6

Сравнение результатов по количеству итераций при выполнении алгоритма *K*-средних показало, что при использовании предложенного метода количество итераций в среднем в 2,9 раза меньше, чем при случайном выборе начальных приближений (табл. 2).

Таблица 2

Количество итераций при различных методах выбора начальных приближений

Название файла	Случайный выбор начальных приближений			Разработанный метод выбора начальных приближений		
	для 25 кластеров	для 50 кластеров	для 75 кластеров	для 25 кластеров	для 50 кластеров	для 75 кластеров
3096	31	39	42	11	17	10
26031	40	17	29	13	8	7
35010	62	41	67	23	17	22
35091	28	35	39	15	18	12
41004	19	32	24	10	15	16
41025	36	27	31	11	7	9
54005	24	26	17	7	7	7
78004	57	22	22	11	9	12
86000	52	41	106	21	20	56
144067	37	39	43	22	13	12

Для сравнения результатов по значению ошибки кластеризации были рассчитаны значения параметров MSE и PSNR. Анализ результатов показал, что в 88 % случаев значения MSE при использовании предложенного метода меньше, чем при случайном выборе начальных приближений. Соответственно уровень PSNR в 88 % случаев выше, чем при случайном выборе начальных приближений. MSE при использовании предложенного метода составляет в среднем 54,93 при кластеризации на 25 кластеров, 58,7 – на 50 кластеров, 31,7 – на 75 кластеров, что соответственно на 3,77, 1,97 и 1,94 в среднем меньше значений MSE при случайном выборе начальных приближений (табл. 3).

Таблица 3

Значения MSE при различных методах выбора начальных приближений

Название файла	Случайный выбор начальных приближений			Разработанный метод выбора начальных приближений		
	для 25 кластеров	для 50 кластеров	для 75 кластеров	для 25 кластеров	для 50 кластеров	для 75 кластеров
3096	37,060	21,729	17,393	33,517	19,492	14,228
26031	33,073	18,935	14,017	32,803	18,331	13,501
35010	55,960	33,462	23,414	53,859	31,855	23,622
35091	64,042	38,292	29,784	62,437	35,718	26,714
41004	21,046	11,404	7,908	18,530	10,289	7,294
41025	17,770	10,184	8,509	15,970	8,961	6,663
54005	22,274	11,735	8,263	18,656	10,968	7,908
78004	32,665	18,288	11,924	30,545	17,070	11,853
86000	110,533	60,187	37,875	89,479	54,016	41,068
144067	56,752	29,273	22,884	51,702	29,122	20,708

В свою очередь, PSNR при использовании предложенного метода составляет в среднем 33,46 дБ при кластеризации на 25 кластеров, 33,85 дБ – на 50 кластеров, 35,17 дБ – на 75 кластеров, что соответственно на 0,34, 0,32 и 0,37 дБ в среднем выше значений PSNR при случайном выборе начальных приближений (табл. 4).

Таблица 4

Значения PSNR при различных методах выбора начальных приближений, дБ

Название файла	Случайный выбор начальных приближений			Разработанный метод выбора начальных приближений		
	для 25 кластеров	для 50 кластеров	для 75 кластеров	для 25 кластеров	для 50 кластеров	для 75 кластеров
3096	36,029	41,630	43,026	40,623	43,999	45,435
26031	34,673	37,368	37,603	35,104	37,577	38,815
35010	30,652	32,885	34,436	30,818	33,099	34,398
35091	30,066	32,300	33,391	30,176	32,602	33,863
41004	34,899	37,560	39,150	35,452	38,007	39,501
41025	35,634	38,052	38,832	36,098	38,607	39,894
54005	34,653	37,436	38,959	35,423	37,729	39,150
78004	32,990	35,509	37,367	33,281	35,808	37,392
86000	27,696	30,336	32,347	28,614	30,806	31,996
144067	30,591	33,466	34,536	30,996	33,489	34,969

Заключение

Разработанный метод выбора начальных приближений позволяет сократить количество итераций алгоритма *K*-средних в 2,9 раза по сравнению с существующими методами. Нулевые кластеры в ходе обработки данных отсутствуют. По сравнению с другими методами ошибка кластеризации контролируется на стадии выбора начальных приближений, а за счет того, что полученные начальные приближения заменяются ближайшими элементами исходного набора данных, она еще уменьшается. После выполнения *K*-средних в 88 % случаев значения MSE при использовании предложенного метода меньше, чем при случайном выборе начальных приближений. Соответственно уровень PSNR в 88 % случаев выше, чем при случайном выборе начальных приближений.

Разработанный метод выбора начальных приближений центров кластеров для алгоритма *K*-средних может успешно применяться в цифровой обработке изображений (задачах кластеризации, сегментации, квантования цветов) и статистическом анализе данных.

Список литературы

1. MacQueen, J.B. Some methods for classification and analysis of multivariate observations / J.B. MacQueen // Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. – Berkley : University of California Press, 1967. – Vol. 1. – P. 281–297.

2. Bradley, P.S. Refining initial points for k-means clustering / P.S. Bradley, U.M. Fayyad // Proceedings 15th International Conf. on Machine Learning. – San Francisco, CA, 1998. – P. 91–99.

3. Pena, J. An Empirical comparison of four initialization methods for the k-means algorithm / J. Pena, J. Lozano, P. Larranaga // Pattern Recognition Letters. – 1999. – Vol. 20. – P. 1027–1040.

4. Likas, A. The Global k-means Clustering algorithm / A. Likas, N. Vlassis, J.J. Verbeek // Pattern Recognition. – 2003. – Vol. 36. – P. 451–461.

5. Khan, S.S. Cluster Center Initialization for K-mean Clustering / S.S. Khan, A. Ahmad // Pattern Recognition Letters. – 2004. – Vol. 25. – P. 1293–1302.

6. Al-Daoud, M.B. A New Algorithm for Cluster Initialization / M.B. Al-Daoud // Proceedings of World Academy of science, engineering and technology. – 2005. – Vol. 4. – P. 74–76.

7. Deelers, S. Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance / S. Deelers, S. Auwatanamongkol // Proceedings of World Academy of science, engineering and technology. – 2007. – Vol. 26. – P. 323–328.

8. Babu, G. A near optimal initial seed value selection in k-means algorithm using a genetic algorithm / G. Babu, M. Murty // Pattern Recognition Letters. – 1993. – Vol. 14. – P. 763–769.

9. Berkeley Segmentation Dataset [Electronic resource]. – Mode of access : http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench. – Date of access : 05.10.2009.

Поступила 13.08.09

Одесская национальная академия
связи им. А.С. Попова,
Украина, Одесса, Кузнечная, 1
e-mail: vampiter@rambler.ru,
kumish@mail.ru

V.I. Zagrebnuk, V.U. Kumish

CLUSTER CENTER INITIALIZATION METHOD FOR *K*-MEANS ALGORITHM

Cluster center initialization method for *K*-means algorithm is developed. The method allows reducing the number of iterations of *K*-means algorithm for 2.9 times in comparison with the existing methods. There are no empty clusters when processing the data. The computational complexity of the method is lower compared to existing ones. This is because the *K*-means is not used to find the initial cluster centers. The clustering error is controlled at the stage of cluster center initialization. Also, the clustering accuracy is improved due to the use of all attribute values for cluster center initialization and replacing the derived values of cluster centers with the nearest elements from the data set.