ДОКЛАСТЕРИЗАЦИЯ КАК СПОСОБ ОПТИМИЗАЦИИ ВРЕМЕНИ АНАЛИЗА ИСХОДНЫХ ДАННЫХ

И. М. Нейский

(МГТУ им. Н.Э. Баумана, г. Москва)

В статье описывается кластерный анализ, его задача, возможности его практического применения. В статье рассматривается метод адаптивной кластеризации ADAKL, описываются его входные параметры, этапы обработки исходных данных и класс задач, для которого предназначен метод. Для описанного метода в статье приводится общая аналитическая оценка сложности. В случае необходимости расширения исходных данных в процессе исследования набора исходные данных в статье предлагается метод докластеризации, приводится его описание, а также описание практических исследований.

Большинство современных предприятий используют в своей деятельности информационные системы. Хранилища данных, в которых собираются данные по бизнес – процессам компании. Объемы накапливаемой информации увеличиваются с течением времени, поэтому актуальной задачей в развитии компании является переход от анализа тенденций текущих показателей деятельности предприятия к более комплексном подходу «извлечения знаний» из имеющихся данных в целях выявления закономерностей.

Изучением проблем и созданием решений в этой области активно занимаются направления Business Intelligence (Интеллектуальный анализ данных) и Knowledge Management (Управление знаниями), в рамках которых выделяются поднаправления Knowledge Discovery in Databases (Выявление знаний в базах данных), Data Mining (Анализ фактографических данных), Text Mining (Анализ неструктурированных данных) и др. Результаты исследований этих направлений положены в основу многих информационно-аналитических систем, которые используются, в основном, для персональной работы экспертов. Однако, современной тенденцией является применение указанных технологий и для централизованного управления организациями.

Кластерный анализ (Cluster analysis или Data Clustering) является одним из этапов интеллектуального анализа данных. Кластеризация, используя свободный поиск, выделяет в данных признаки, по которым данные можно поделить на группы. Процесс кластеризации неоднозначен, поскольку группировка данных целиком зависит от способа, по которому измеряется информационное расстояние между записями набора данных.

Примерами практических задач, в которых используется, требуется или планируется применение кластерного анализа, являются следующие задачи:

- выделение групп клиентов брокерского обслуживания для формирования перечня предлагаемых сервисов;
- формирование потребительской корзины;
- принятие решения о выдаче потребительского кредита;
- сегментирование сферы деятельности с целью повышения эффективности производительности;
- обработка изображений;
- тематический анализ библиотеки документов;
- оптимизация использования складских помещений;

- выявление транзакций, проведенных по поддельным кредитным картам;
- выявление потенциальных болезней пациентов;
- построение показательной (репрезентативной) выборки и т.д.

На данный момент известно более 50 методов кластеризации, среди которых довольно большое количество методов представлено в математической, алгоритмической форме, но значительно меньше методов имеют реализацию и рекомендацию по области применения алгоритма. Знание того, какие методы дают наилучший результат, может подсказать направление движения тем, кто создаёт новые алгоритмы или совершенствует существующие.

Количество объектов исследований с целью разбиения на кластеры постоянно растет, в том числе во время проведения анализа сформированного массива фактографических данных, увеличивая время проведения исследований, поэтому возникает задача оптимизации времени анализа исходных данных в виде докластеризации «новых» объектов. В общем случае, при увеличении количества исследуемых объектов, требуется повторный запуск исследования на всем массиве данных, который потребует соответствующих затрат времени, технических и человеческих ресурсов. Задачей докластеризации является определение необходимости повторного запуска исследования полного массива данных и, в случае отсутствия признаков появления новых значимых групп объектов, распределение «новых» объектов по имеющимся кластерам на основе оценки близости распределяемых объектов к распределенным объектам.

В рамках данной статьи рассматривается метод адаптивной кластеризации ADAKL, предназначенный для решения класса задач, характеризующегося следующими характеристиками:

- объекты кластеризации – более 10 000;
- параметры объекта кластеризации – более 20;
- время выполнения кластеризации – не более 5 часов;
- количество получаемых кластеров – не более 15 – 20;
- параметры объектов кластеризации: числовые и текстовые.

В основе данного метода используются методы MST [2] и Fuzzy C-means [3].

За счет использования принципов иерархического метода MST разработанный метод на начальном этапе получает оптимизированную древовидную структуру и первичные кластерные центры, которые уточняются во второй части исследования. Двухэтапность выполнения кластеризации и использование оценочной функции разбиения позволяет повысить качество проводимой кластеризации. Вычисление глобального критерия делает метод кластеризации во много раз быстрее, чем при использовании локального критерия при парном сравнении объектов, поэтому «глобализация» оценочной функции – один из путей получения масштабируемых алгоритмов. Использование нечеткости при определении объектов по кластерам позволяет сделать более полное разбиение исходного множества на кластеры, ликвидируя тем самым неопределенности, которые возникают при четком разбиении.

Описание метода «ADAKL»:

Входные данные метода:

$D = \{u_1, u_2, ..., u_m\}$, где u_i – объекты кластеризации, m – количество объектов кластеризации, $i = \overline{1, m}$;

$u_i = \{(Value_{i1}, t_1), (Value_{i2}, t_2), ..., (Value_{in}, t_n)\}$, где $Value_{ij}$ – значение $j^{\text{ой}}$ атрибута $i^{\text{ой}}$ объекта кластеризации, t_j – тип атрибута объекта кластеризации, n – количество атрибутов объекта кластеризации, $j = \overline{1, n}$;

$t_j = \{ValueType_j, FieldType_j\}$, где $ValueType_j$ – тип значения атрибута, $ValueType_j \in ValueTypes$, $FieldType_j$ – вид значения атрибута, $FieldType_j \in FieldTypes$;

Множество типов значений атрибутов:

$$ValueTypes = \begin{Bmatrix} Öåëî ÷èñëåí íûé òèï, \\ Äåí åæ íûé òèï, \\ Ëèí åéíîñòè òèïé \end{Bmatrix},$$ где

$Öåëî ÷èñëåí íûé òèï \subset \mathbf{Z}$, $Äåí åæ íûé òèï \subset R$, $Ëèí åéíîñòè òèïé \subset Ñëî âàðí àÿ ñèñòå ìà$;

$$Ñëî âàðí àÿ ñèñòå ìà = \begin{Bmatrix} Ëèí êâ.òèï1, \\ Ëèí êâ.òèï2, \\ ....., \\ Ëèí êâ.òèïS \end{Bmatrix}, \text{где}$$

$Ëèí êâ.òèï_i$ – объект словарной системы, характеризующий оценочные/качественные показатели объектов кластеризации;

Множество видов значений атрибутов:

$$FieldTypes = \begin{Bmatrix} Âõîä íîå, \\ Èäåíòèôèêàòîðíûé ñòîëáåö, \\ Íåèíôîðìàòèâíîå \end{Bmatrix},$$ где

«$Âõîä íîå$» – означает участие атрибута объекта в методе, «$Èäåíòèôèêàòîðíûé ñòîëáåö$» – обозначает ключевой атрибут объектов кластеризации, идентифицирующий каждый объект входного набора данных, «$Íåèíôîðìàòèâíîå$» – обозначает атрибут объекта, не оказывающий влияние на результаты работы метода.

q – максимальное количество кластеров, $q \le m$;

$K = \{K_1, K_2, ..., K_n\}$, где K_i – весовой коэффициент влияния атрибута объекта, $K_i \in [0;1]$;

p – размазанность кластеров, $p \in (0;10]$;

w – степень удаленности элементов, $w \in (0;1]$;

$Metric$ – способ определения расстояния между объектами, $Metric \in Metrics$;

Множество способов определения расстояния между объектами:

$$Metrics = \begin{Bmatrix} Âåëëèäîâ ðàññò îÿíèå, Êâàäðàò Âåëëèäîâ ðàññò îÿíèÿ, ðàññò îÿíèå×åêîâ ø åâà \end{Bmatrix};$$

$OstTreeMethod$ – способ построения минимального остовного дерева, $OstTreeMethod \in OstTreeMethods$;

$$OstTreeMethods = \begin{Bmatrix} Àëãîðèòì Áîðóâêè, Àëãîðèòì \\ Êðóñêàëà, Àëãîðèòì Ïðèìà \end{Bmatrix};$$

$NormMethod$ – способ проведения нормализации значений числовых атрибутов, $NormMethod \in NormMethods$;

$$NormMethods = \begin{Bmatrix} Ëèíåéíàÿ íîðìàëèçàöèÿ, \\ Ñòàòèñòè÷åñêàÿ íîðìàëèçàöèÿ \end{Bmatrix}$$

Выходные данные метода:

$$C = \left\{ C_1, C_2, ..., C_c \middle| O^c \to max, c \le q, C_1 \cup C_2 \cup ... \cup C_c = D \right\}$$, $u_i \overset{\mu_{ij} \in [0,1]}{\in} C_j$, $i = \overline{1,m}$, $j = \overline{1,c}$

Тело метода:

Этап 1. Нормализация значений числовых атрибутов.

В случае линейной нормализации выполняется следующее:

$$Value_{ij} := \left\{ \frac{Value_{ij}}{\underset{j}{Max}(Value_{ij})} \middle| \begin{matrix} \underset{j}{Max}(Value_{ij}) \ne 0, \\ t_j \in \{Öåëî ÷èñëåí íûé òèï, Äåí åæ íûé òèï\} \end{matrix} \right\}$$

В случае статистической нормализации выполняется следующее:

$$Value_{ij} := \left\{ \frac{\left| Value_{ij} - \frac{\sum_{i=1}^{m} Value_{ij}}{m} \right|}{\sqrt{\frac{\sum_{i=1}^{m} (Value_{ij})^2}{m} - \left(\frac{\sum_{i=1}^{m} Value_{ij}}{m} \right)^2}} \middle| \begin{matrix} \sqrt{\frac{\sum_{i=1}^{m} (Value_{ij})^2}{m} - \left(\frac{\sum_{i=1}^{m} Value_{ij}}{m} \right)^2} \ne 0, \\ t_j \in \begin{Bmatrix} Öåëî ÷èñëåí íûé òèï, \\ Äåí åæ íûé òèï \end{Bmatrix} \end{matrix} \right\}$$

Этап 2. Вычисление матрицы взаимных расстояний между объектами.

$Dist_{ij} = \|u_i - u_j\| = Metric(u_i, u_j)$, где $Metric$ – способ определения расстояния между объектами.

За основу методик оценки расстояний взяты классические формулы и доработаны для использования в методе:

Если $Metric = Åâêëèäîâ ðàññòîÿíèå$, то $Dist_{ij} = \sqrt{\sum_w \left(\left[Value_{iw} - Value_{jw}\right] * K_w\right)^2}$,

Если $Metric = Êâàäðàò Åâêëèäîâà ðàññòîÿíèÿ$, то $Dist_{ij} = \sum_w \left(\left[Value_{iw} - Value_{jw}\right] * K_w\right)^2$,

Если $Metric = Ðàññòîÿíèå×åáûø åâà$, то $Dist_{ij} = \underset{w}{Max}\left[Value_{iw} - Value_{jw}\right] * K_w$,

где $i, j \in [1, m], w = \overline{1,n}$ при условии $FieldType[w] = "Âõîäíîå"$

Этап 3. Построение минимального остовного дерева.

Построение минимального остовного дерева выполняется по выбранному способу построения дерева с использованием матрицы взаимных расстояний между объектами $Dist$.

Этап 4. Разделение объектов на кластеры и построение матрицы нечеткого разбиения F.

Матрица нечеткого разбиения: $F = \left[\mu_{ij}\right], \mu_{ij} \in [0,1], i \le q, j = \overline{1,m}$, где μ_{ij} – степень принадлежности $i^{ãî}$ объекта к $j^{ì ó}$ кластеру. Матрица разбиения обладает следующими свойствами: $\sum_{i=1}^{k} \mu_{ij} = 1, j = \overline{1,m}$, $0 < \sum_{j=1}^{m} \mu_{ij} \le m, i = \overline{1,k}$.

Шаг 1. Определение количества кластеров итерации: $k := q$.

Шаг 2. Разделение минимального остовного дерева на k кластеров на основании длины ребер дерева по убыванию величины: $Dist_{ij}^k := \left\{0 \middle| Dist_{ij}^{k\ i,j=\overline{1,m}} = Max\right\}$.

Шаг 3. Расчет центров выделенных кластеров V_i^k.

$V_i^k = Avg\left(\left\{u_j \middle| u_j \in C_i^k\right\}\right)$, где Avg – оператор вычисления среднего значения показателей объектов, входящих в кластер k, $i = \overline{1,k}$, $j = \overline{1,m}$.

Для числовых типов оператор Avg определяется выражением:

$$Avg[r] = \frac{\sum_{u_j \in V_i^k} \left\{Value_{jr} \middle| FieldType[w] = "Âõîäíîå"\right\}}{\left|V_i^k\right|},$$

$j = \overline{1,m}$, $r = \overline{1,n}$.

Для лингвистических типов оператор Avg определяется выражением, учитывающим взаимное расстояние между значениями анализируемого атрибута объектов в целях его минимизации или выбирающего значение атрибута, имеющего наибольшую частоту повторяемости:

$$Avg[r] = \left\{ \frac{Value_{jr}}{\sum_{u_j \in V_i^k, u_l \in V_i^k} \left\|Value_{jr} - Value_{lr}\right\| = min\ èëè\ \underset{u_j \in V_i^k}{\varphi = max}} \middle| FieldType[w] = "Âõîäíîå" \right\}$$

, $r = \overline{1,n}$, $j = \overline{1,m}$, $l = \overline{1,m}$, где φ – частота значения атрибута в пределах кластера V_i^k.

Шаг 4. Расчет матрицы расстояний от объектов до центров кластеров V_i^k: $Dist_{ij}^k = \left\|V_i^k - u_j\right\| = Metric\left(V_i^k, u_j\right)$, $i = \overline{1,k}$, $j = \overline{1,m}$, где $Metric$ – способ определения расстояния между объектами.

За основу методик оценки расстояний взяты классические формулы и доработаны для использования в методе:

Если $Metric = Åâêëèäîâ ðàññòîÿíèå$, то $Dist_{ij}^k = \sqrt{\sum_w \left(\left[Value_{iw} - Value_{jw}\right] * K_w\right)^2}$,

Если $Metric = Êâàäðàò Åâêëèäîâà ðàññòîÿíèÿ$, то $Dist_{ij}^k = \sum_w \left(\left[Value_{iw} - Value_{jw}\right] * K_w\right)^2$,

Если $Metric = Ðàññòîÿíèå×åáûø åâà$, то $Dist_{ij}^k = \underset{w}{Max}\left[Value_{iw} - Value_{jw}\right] * K_w$,

где $i, j \in [1, m], w = \overline{1,n}$ при условии $FieldType[w] = "Âõîäíîå"$.

Шаг 5. Нормализация матрицы расстояний от объектов до центров кластеров V_i^k:

$$Dist_{ij}^{k'} = \begin{cases} \dfrac{Dist_{ij}^{k}}{Max\left(Dist_{ij}^{k}\right)}, Max\left(Dist_{ij}^{k}\right) \neq 0 \\ Dist_{ij}^{k}, Max\left(Dist_{ij}^{k}\right) = 0 \end{cases}, \quad i = \overline{1,k},$$

$j = \overline{1,m}$.

Шаг 6. Соотнесение объектов к кластерам в соответствии со степенью удаленности элементов кластера (w):

$$u_j \in V_i^k \left| Dist_{ij}^{k'} \le w \text{ èëè } Dist_{ij}^{k'} = \underset{i}{Min}\left(Dist_{ij}^{k'}\right), \quad i = \overline{1,k}, \right.$$

$j = \overline{1,m}$.

Шаг 7. Расчет степени принадлежности кластеру.

$\mu_{ij} = \left(1 - Dist_{ij}^{k'}\right)^2$, $i = \overline{1,k}$, $j = \overline{1,m}$.

Шаг 8. Нормализация матрицы нечеткого разбиения:

$$\mu_{ij} = \left\{ \begin{array}{l} \dfrac{\mu_{ij}}{\sum\limits_{i=1}^{k} \mu_{ij}} \left| \sum\limits_{i=1}^{k} \mu_{ij} \neq 0 \right. \\ \mu_{ij} \left| \sum\limits_{i=1}^{k} \mu_{ij} = 0 \right. \end{array} \right\}, \quad j = \overline{1,m}.$$

Шаг 9. Вычисление центров полученных кластеров с использованием матрицы нечеткого разбиения: $V_i^{k'} = \dfrac{\sum\limits_{j=1}^{m} \mu_{ij}^{\ p} * u_j}{\sum\limits_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} \mu_{ij}^{\ p}}, i = \overline{1,k}$.

Для лингвистических атрибутов центра кластера вычисление производится с использованием выражения: $V_i^{k'}[r] = \underset{\mu_{ij} = Max\left(\mu_{lj}\right)}{Value_{jr}}$.

Шаг 10. Оценка качества полученного разбиения.

Оценка качества полученного разбиения на k кластеров с использованием полученных центров кластеров:

$$O^k = \frac{\sum\limits_{i=1,k} \dfrac{\left|V_i^{k'}\right| * \sum\limits_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} \mu^{p}_{\ ij} * \left\|V_i^{k'} - u_j\right\|}{\underset{i \neq j}{\underset{u_j \in V_i^{k'}}{Min}}\left(\left\|V_i^{k'} - u_j\right\|\right) * Ma \underset{u_j \in V_i^{k'}}{\left(\left\|V_i^{k'} - u_j\right\|\right)} * \sum\limits_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} \left\|V_i^{k'} - u_j\right\| * k}}{m * k^2}, \text{ где}$$

$\left|V_i^{k'}\right|$ – количество элементов в кластере i;

$\left\|V_i^{k'} - u_j\right\| = Metric\left(V_i^{k'}, u_j\right)$ – расстояние от центра кластера i до элемента u_j;

$u_j \in V_i^{k'}$ – отражение условия о принадлежности элемента кластеру.

Шаг 11. $k := k - 1$.

Шаг 12. Если $k > 0$, то переход на шаг 2.

Этап 5. Выбор наилучшего разбиения: $O_{\hat{I}\,\ddot{i}\hat{o}} = \underset{i=1,q}{MAX}\left(O^i\right)$.

Общая аналитическая оценка сложности метода является комплексной ввиду наличия его этапного разделения. Для оценки аналитической сложности метода введем следующие обозначения:

a – количество входных числовых атрибутов;

b – количество входных лингвистических атрибутов;

m – количество кластеризуемых объектов;

k – текущее количество кластеров;

q – общее количество кластеров.

Общая оценка метода:

Для случая с линейной нормализацией и использованием алгоритма Борувки:

$O(2*a*m) + O\left(m^2 * [a+b]\right) + O\left(m^2 \cdot Lg(m)\right) + O\left(m^2 * (a+b+q)\right) + O(q) \equiv$
$O\left(a*m + m^2 * [a+b] + m^2 \cdot Lg(m) + m^2 * (a+b+q) + q\right) \equiv$
$O\left(m^2 * \left(a + b + Lg(m) + q\right)\right)$

Для случая с линейной нормализацией и использованием алгоритма Крускала:

$O(2*a*m) + O\left(m^2 * [a+b]\right) + O\left(m^2 \cdot Lg\left(m^2\right)\right) + O\left(m^2 * (a+b+q)\right) + O(q) \equiv$
$O\left(a*m + m^2 * [a+b] + m^2 \cdot Lg\left(m^2\right) + m^2 * (a+b+q) + q\right) \equiv$
$O\left(m^2 * \left(a + b + Lg\left(m^2\right) + q\right)\right)$

Для случая с линейной нормализацией и использованием алгоритма Прима:

$O(2*a*m) + O\left(m^2 * [a+b]\right) + O\left(m^2 \cdot Lg(m)\right) + O\left(m^2 * (a+b+q)\right) + O(q) \equiv$
$O\left(a*m + m^2 * [a+b] + m^2 \cdot Lg(m) + m^2 * (a+b+q) + q\right) \equiv$
$O\left(m^2 * \left(a + b + Lg(m) + q\right)\right)$

Для случая со статистической нормализацией и использованием алгоритма Борувки:

$O\left(3*a*m^2\right) + O\left(m^2 * [a+b]\right) + O\left(m^2 \cdot Lg(m)\right) + O\left(m^2 * (a+b+q)\right) + O(q) \equiv$
$O\left(a*m^2 + m^2 * [a+b] + m^2 \cdot Lg(m) + m^2 * (a+b+q) + q\right) \equiv$
$O\left(m^2 * \left(a + b + Lg(m) + q\right)\right)$

Для случая со статистической нормализацией и использованием алгоритма Крускала:

$$O\left(3*a*m^2\right)+O\left(m^2*[a+b]\right)+O\left(m^2\cdot Lg\left(m^2\right)\right)+O\left(m^2*(a+b+q)\right)+O(q)\equiv$$
$$O\left(a*m^2+m^2*[a+b]+m^2\cdot Lg\left(m^2\right)+m^2*(a+b+q)+q\right)\equiv$$
$$O\left(m^2*\left(a+b+Lg\left(m^2\right)+q\right)\right)$$

Для случая со статистической нормализацией и использованием алгоритма Прима:

$$O\left(3*a*m^2\right)+O\left(m^2*[a+b]\right)+O\left(m^2\cdot Lg\left(m\right)\right)+O\left(m^2*(a+b+q)\right)+O(q)\equiv$$
$$O\left(a*m^2+m^2*[a+b]+m^2\cdot Lg\left(m\right)+m^2*(a+b+q)+q\right)\equiv$$
$$O\left(m^2*\left(a+b+Lg\left(m\right)+q\right)\right)$$

Итоговые оценки метода:

$O\left(m^2*\left(a+b+Lg\left(m\right)+q\right)\right)$, $O\left(m^2*\left(a+b+Lg\left(m^2\right)+q\right)\right)$.

Для расширения исходных данных в процессе проведения анализа необходимо произвести дополнительное исследование добавляемых данных, расширив входные параметры основного метода параметром δ – доверительный интервал, и расширить метод проведения анализа исходных данных ADAKL методом докластеризации (рис. 1).

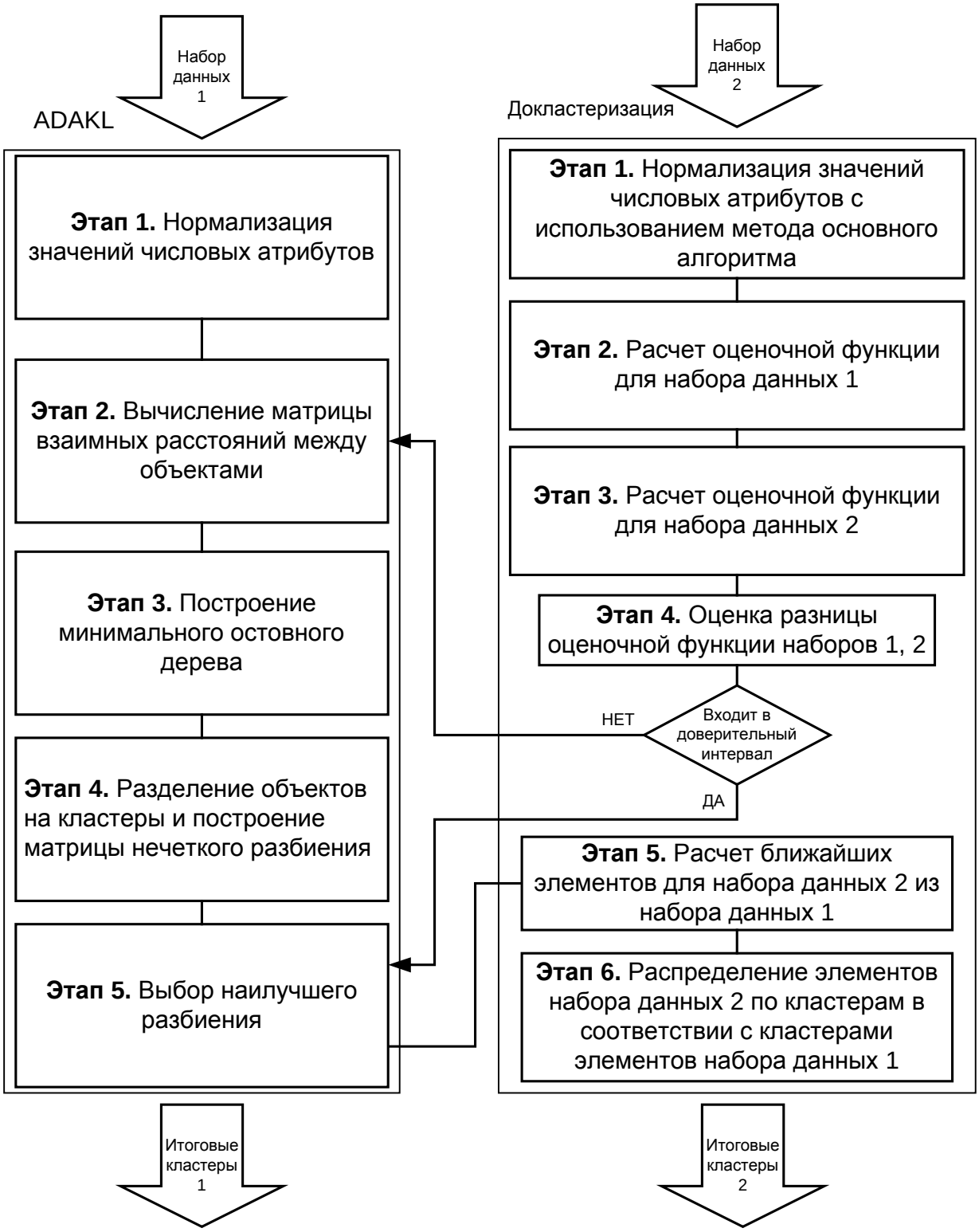

Рис. 1. Расширение исходного набора данных в процессе исследования.

Расчет оценочной функции в данном случае осуществляется следующим образом:

$$O_1 = \sqrt{\frac{\sum_{i=1}^{r} \left\| A_i - Avg[A] \right\|^2}{r}}, \quad O_2 = \sqrt{\frac{\sum_{i=1}^{o} \left\| B_i - Avg[B] \right\|^2}{o}}, \quad (1)$$

где

O_1, O_2 – оценочная функция исходного набора данных 1 и 2 соответственно;

A, B – исходные наборы данных 1 и 2 соответственно;

r, o – количество объектов в исходных наборах данных 1 и 2 соответственно;

$\|A_i - Avg[A]\|, \|B_i - Avg[B]\|$ – оператор вычисления расстояния между объектом и средним значением множества, полученного с использованием оператора вычисления среднего значения основного метода (ADAKL), для исходных наборов данных 1 и 2 соответственно. Вычисление данного расстояния выполняется с учетом весовых коэффициентов основного метода: $K = \{K_1, K_2, ..., K_n\}$, где K_i – весовой коэффициент влияния атрибута объекта, $K_i \in [0;1]$.

Оценка разницы оценочных функций на этапе 4 выполняется следующим образом: $|O_1 - O_2| \le \delta$, где δ – доверительный интервал.

Выявление ближайших элементов из множества A для множества B выполняется с помощью метода, используемого в ADAKL, и определяется на этапе конфигурирования метода.

Распределение элементов множества B по вычисленному расстоянию до ближайших k – объектов[1] из множества A выполняется следующим образом:

$\mu_{ij} = \left(1 - Dist_k^{Norm}\right) * \mu_{kj} = \left[1 - Dist_k^j / Max\left(Dist_k\right)\right] * \mu_{kj}$, где

$i = \overline{1, o}$ – порядковый номер элемента из множества B;

$j \in [1, r]$ – порядковый номер ближайшего элемента из множества A для соответствующего элемента из множества B;

k – порядковый номер ближайшего элемента;

$Dist_k^j = \|B_i - A_j\|$ – расстояние между ближайшими элементами из множеств A и B;

$Max\left(Dist_k\right)$ – максимальное расстояние от элемента из множества B до элемента из множества A;

μ_{kj} – степень принадлежности элемента из множества A к кластеру k.

Рассмотренный метод «ADAKL» обладает следующими достоинствами:

- двухэтапная кластеризация фактографических данных;
- способен работать с лингвистическими атрибутами объектов кластеризации с применением нечеткой логики и введением словарной системы для вычисления расстояний между объектами входного набора данных;
- использует весовые коэффициенты для анализируемых атрибутов объектов с целью повышения/понижения влияния атрибутов на результаты кластеризации и адаптации метода к различным предметным областям;
- использует степень удаленности объектов/элементов для соотнесения объектов в кластеры при разделении;
- использует размазанность кластера, для определения нечеткости отнесения объекта к кластеру;
- использует способ определения расстояния между объектами, разработанный на основе базовых метрик: Евклидово расстояние, Квадрат Евклидова расстояния, расстояние Чебышева, с введением в функцию вычисления расстояния весовых коэффициентов и логики по вычислению расстояний между значениями лингвистических атрибутов;
- предлагает три способа построения минимального остовного дерева, результат работы которых одинаков, но все три способа отличаются разной вычислительной сложностью, что является определяющим на больших объемах данных: алгоритм Борувки – $O\left(\|Dist\|^2 \cdot Lg(m)\right)$, алгоритм

[1] k – оптимальное количество кластеров в соответствии с критерием: $O_{îïò} = \underset{i=1,q}{MAX}\left(O^i\right)$

Крускала – $O\left(\|Dist\|^2 \cdot Lg\|Dist\|\right)$, алгоритм Прима – $O\left(\|Dist\|^2 \cdot Lg(m)\right)$;

- использует критерий оценки разбиения на кластеры с учетом специфики предметной области: небольшое количество кластеров, наибольшая плотность, средняя удаленность объектов;
- способен выполнить кластеризацию дополнительного набора данных в процессе исследования основного набора данных с полным применением метода, начиная с этапа 2, или докластеризовать набор данных с учетом полученных результатов на этапе 5.

Рассмотренный метод «ADAKL» обладает следующими недостатками:

- квадратичная зависимость аналитической сложности метода от количества исходных данных по объектам кластеризации.

Использование в методе принципа докластеризации позволяет достичь уменьшения времени исследования за счет нахождения взаимосвязей между исследованными объектами и добавляемыми объектами. Данное решение имеет ограничение в виде оценки обоих множеств по предложенному выше критерию (1), которое определяет возможность объединения множеств без повторения полного исследования исходных наборов данных. Для анализа времени работы метода[2] с использованием докластеризации проведены практические исследования 30 наборов данных с разным количеством записей и типами исследуемых атрибутов объектов. Для исследования исходных наборов данных использовался метод кластерного анализа «ADAKL» и его расширяющий метод докластеризации (табл. 1).

[2] Временные характеристики работы получены с использованием персонального компьютера модели Fujitsu-Siemens AMILO Xi 1546.

Показатель / Эксперимент	Количество записей во входном наборе данных (шт.)	Этап 1 основного метода (мс.)	Этап 2 основного метода (мс.)	Этап 3 основного метода (мс.)	Этап 4 основного метода (мс.)	Этап 5 основного метода (мс.)	Общее время выполнения основного метода (мс.)	Время выполнения докластеризации (мс.)
Общее количество атрибутов – 74, количество числовых атрибутов – 8, количество лингвистических атрибутов – 0, исходное количество кластеров – 30								
1	50	1	1	15	484	1	502	1
2	100	1	16	218	671	15	921	32
3	150	16	15	1 110	859	1	2 001	31
4	200	15	47	3 516	1 077	1	4 656	46
5	250	1	78	8 640	1 312	1	10 032	78
6	300	1	108	18 470	1 686	15	20 280	108
7	350	1	171	34 408	1 968	1	36 549	156
8	400	1	219	59 940	2 265	16	62 441	218
9	450	1	281	95 161	2 593	16	98 052	266
10	500	1	329	146 365	2 968	15	149 678	328
Общее количество атрибутов – 74, количество числовых атрибутов – 39, количество лингвистических атрибутов – 0, исходное количество кластеров – 30								
11	50	1	15	15	922	1	954	15
12	100	1	46	218	1 531	1	1 797	47
13	150	1	108	1 108	2 172	1	3 390	109
14	200	1	188	3 562	3 077	16	6 844	188
15	250	15	282	8 764	3 531	1	12 593	295
16	300	0	421	18 172	4 218	1	22 812	421
17	350	1	578	34 343	4 921	1	39 844	577
18	400	1	749	58 080	5 672	1	64 503	719
19	450	1	952	93 877	6 438	15	101 283	937
20	500	1	1 171	144 988	7 202	15	153 377	1 156
Общее количество атрибутов – 5, количество числовых атрибутов – 4, количество лингвистических атрибутов – 1, исходное количество кластеров – 9								
21	50	1	1	15	2 219	1	2 237	1
22	100	1	1	171	4 015	1	4 189	1
23	150	1	15	859	5 954	1	6 830	15
24	200	15	15	2 688	8 046	1	10 765	15
25	274	1	31	6 609	9 751	1	16 393	30
26	348	1	46	13 001	11 704	1	24 753	46
27	422	1	61	22 674	13 610	1	36 347	62
28	496	1	78	39 955	15 532	1	55 567	78
29	570	1	94	69 160	21 079	1	90 335	94
30	644	1	139	117 024	22 158	1	139 323	139

Таблица 1. Данные и результаты практических исследований.

Анализ временных характеристик исследований исходных наборов данных с помощью «ADAKL» выявил следующее:

- средневзвешенное время анализа объекта, равное отношению общего времени выполнения анализа к количеству записей во входном наборе данных, увеличивается с возрастанием количества объектов и количества атрибутов. Степень изменения средневзвешенного времени анализа при изменении количества объектов и количества атрибутов различается (см. рис. 2, 3), а также зависит от характеристик исходного набора данных;
- время, затрачиваемое на одну запись дополнительного набора данных при выполнении

докластеризации, сравнимо с временем, затрачиваемым основным методом на втором этапе при вычислении матрицы взаимных расстояний между объектами. Время выполнения второго этапа по отношению к общему времени выполнения кластеризации составляет приблизительно в среднем 0.008 часть.

Представим затраты времени на работу методов при проведении исследований в графическом виде (рис. 2, 3, 4).

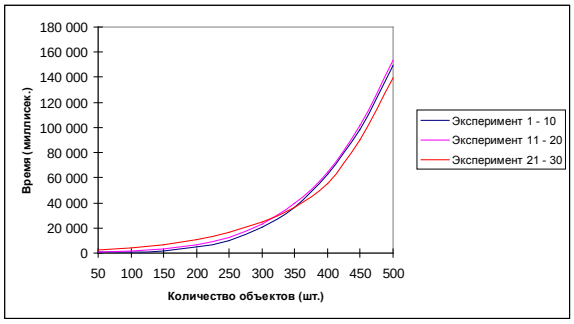

Рис. 2. Зависимость общего времени выполнения метода от количества обрабатываемых объектов.

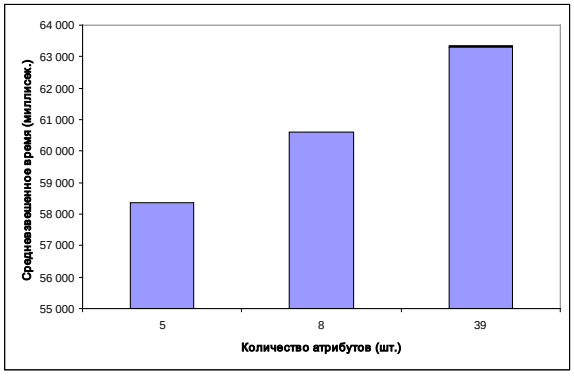

Рис. 3. Зависимость средневзвешенного времени обработки одной записи от количества объектов.

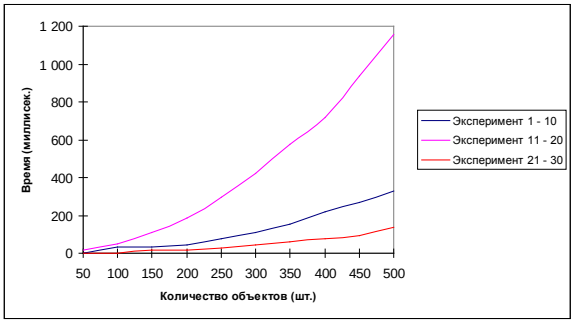

Рис. 4. Зависимость общего времени выполнения докластеризации от количества обрабатываемых объектов.

Проведенные исследования показали, что использование докластеризации при полной или частичной схожести исследуемых наборов данных позволяет добиться значительного уменьшения времени анализа набора данных, не потеряв при этом качественной характеристики проводимого кластерного анализа в связи с использованием аналогичных основному методу методов оценки информационных расстояний между объектами.

Более подробная информация о методе и проведенных исследованиях представлена по адресу http://philippovich.ru/Persons/Neyskiy/Neyskiy.htm.

1. Growe, Glenn A. Comparing Algorithms and Clustering Data: Components of the Data Mining Process [Электронный ресурс] / Glenn A. Growe – Электрон. дан. – [1999]. – Режим доступа: www.gvsu.edu, свободный. – Загл. с экрана. – Яз. англ.

2. He, H. Efficient Algorithms for Mining Significant Substructures in Graphs with Quality Guarantees [Электронный ресурс] / H. He, A.K. Singh // Seventh IEEE International Conference On Data Mining – Электрон. дан. – [2007]. – Режим доступа: www.vldb.org, свободный. – Загл. с экрана. – Яз. англ.

3. Штовба, С.Д. Введение в теорию нечетких множеств и нечеткую логику [Электронный ресурс] / С.Д. Штовба. – Электрон. дан. – [2004]. – Режим доступа: matlab.exponenta.ru, свободный. – Загл. с экрана. – Яз. рус.

4. Кормен, Томас Х., Лейзерсон, Чарльз И., Ривест, Рональд Л., Штайн, Клиффорд Алгоритмы: построение и анализ. Пер. с англ. / Томас Х. Кормен, Чарльз И. Лейзерсон, Рональд Л. Ривест, Клиффорд Штайн. – М.: Издательский дом "Вильямс", 2005. – 2-ое издание.