© 2022 г. Д.А. ЕГУРНОВ (egurnovd@yandex.ru),
Д.И. ИГНАТОВ, канд. техн. наук (dignatov@hse.ru)
(Национальный исследовательский университет
"Высшая школа экономики", Москва)

ТРИКЛАСТЕРЫ БЛИЗКИХ ЗНАЧЕНИЙ ДЛЯ АНАЛИЗА ТРЕХМЕРНЫХ ДАННЫХ[1]

Работа посвящена проблеме трикластеризации в многозначных триадических контекстах в терминах одного из многомерных расширений анализа формальных понятий, которая может быть рассмотрена как поиск плотных подтензоров в трехмерных тензорах над полем действительных чисел. Предлагаются два метода решения этой задачи: NOAC — вариант метода OAC-трикластеризации для числовых данных на основе дельта-операторов и триадическая версия метода k-средних с уточненной метрикой на основе манхэттенского расстояния и предикатов близости по каждому из трех измерений. Проведены численные эксперименты как на реальных, так и на синтетических данных, подтверждающие превосходство метода NOAC в смысле критериев качества найденных трикластеров.

Ключевые слова: трикластеризация, анализ формальных понятий, трехмерные тензоры, многозначные контексты.

DOI: 10.31857/S0005231022060071, **EDN:** ACVQYG

1. Введение

В современном мире характерно наличие насыщенной информационной среды, в которой скорость появления ресурсов и генерации новых данных растет с каждым годом. Часто главной проблемой, препятствующей успешному сбору, систематизации и извлечению знаний из этих данных, является отсутствие в них четко определенной структуры. Проблема анализа данных из неструктурированных источников занимает значительное место в современной прикладной математике и информатике [1]. Кластерный анализ предоставляет широкий спектр подходов и методов для определения внутренней структуры данных и классификации объектов. Основная идея заключается в объединении сходных по некоторым критериям объектов в группы, называемые кластерами.

Существуют методы, решающие задачи кластеризации объектов или признаков (одномерная кластеризация), и альтернативные методы, сохраняющие

[1] Статья подготовлена в результате проведения исследования в рамках Программы фундаментальных исследований Национального исследовательского университета "Высшая школа экономики" (НИУ ВШЭ).

объектно-признаковое описание сходства (бикластеризация) [2]. Тем не менее трикластеризация и n-мерная кластеризация как дальнейшее ее расширение не были столь же тщательно изучены (см., например, обзор [3]). В данной работе авторы ставят целью разработку и исследование алгоритмов для трикластеризации вещественных данных при наличии пропущенных значений, основываясь на методах OAC-трикластеризации (OAC от Object, Attribute, Condition) [4], сравнение эффективности работы этих алгоритмов, а также обработку ими реальных данных и интерпретацию результатов.

В ходе разработки новых алгоритмов были рассмотрены существующие методы, решающие близкие задачи, а именно: OAC-трикластеризация, основанная на штрих-операторах [4], метод шкалирования формальных понятий (Conceptual Scaling) [5] и его реализация TriMax [5], а также метод межпорядкового шкалирования (Interordinal Scaling) [5]. По различным причинам они не могли быть напрямую использованы для решения поставленной в данной работе задачи, поэтому авторы предлагают два метода для поиска трикластеров близких значений в многозначных триадических контекстах: метод NOAC (Numerical OAC), являющийся расширением на многозначный случай OAC-трикластеризации, основанной на штрих-операторах, и классический алгоритм кластеризации k-средних (k-means), использующий авторскую метрику для вычисления расстояний (Tri-k-means).

Данная работа состоит из четырех частей. Первая часть содержит некоторые базовые теоретические определения, формирующие основу предметной области и закладывающие математический аппарат для последующих исследований. В ней также содержатся все общие формулировки, необходимые для понимания описанных в данной работе алгоритмов. Во второй части рассматриваются существующие методы поиска би- и трикластеров в полиадических данных, а также приводится анализ их возможностей. В третьей части предложены два алгоритма, решающие поставленную выше задачу. В четвертой части кратко описаны результаты экспериментов на синтетических и реальных данных. В заключении подводится итог проделанной работе.

2. Базовые определения

2.1. Диадический случай

Для полноценного погружения в предметную область следует начать с формулировки базовых определений анализа формальных понятий (Formal Concept Analysis — FCA) [6].

Пусть даны два множества: G и M. Элементы множества G называют объектами (от нем. Gegenständ – объект), а элементы множества M – признаками (от нем. Merkmal – признак). Пусть также дано бинарное отношение I, являющееся подмножеством декартового произведения этих множеств: $I \subseteq G \times M$. Формальным контекстом называется тройка $K = (G, M, I)$. Если пара (g, m), где $g \in G$ и $m \in M$, принадлежит отношению инцидентности I, говорят, что "объект g обладает признаком m". Это обозначается как $(g, m) \in I$ или gIm.

Обычно формальные контексты представляются в виде матриц или таблиц с булевыми значениями.

Теперь рассмотрим два отображения $\phi : 2^G \to 2^M$ и $\psi : 2^M \to 2^G$. Пусть $\phi(A) = \{m \mid \forall g \in A : gIm\}$, $\psi(B) = \{g \mid \forall m \in B : gIm\}$. Для них выполняются следующие условия: (для $A_1, A_2 \subseteq G; B_1, B_2 \subseteq M$)

$$A_1 \subseteq A_2 \Rightarrow \phi(A_2) \subseteq \phi(A_1),$$
$$B_1 \subseteq B_2 \Rightarrow \psi(B_2) \subseteq \psi(B_1).$$

Эти отображения задают операторы Галуа для множеств объектов и признаков. Так как по составу множества можно однозначно определить, какой из операторов применяется, обычно используется единое обозначение $(\cdot)'$. Верны следующие свойства (для $A, A_1, A_2 \subseteq G, B \subseteq M$):

1) $A_1 \subseteq A_2 \Rightarrow A_2' \subseteq A_1'$,
2) $A_1 \subseteq A_2 \Rightarrow A_1'' \subseteq A_2''$,
3) $A \subseteq A''$,
4) $A' = A'''$ (также $A'' = A''''$),
5) $(A_1 \cup A_2)' \Leftrightarrow A_1' \cap A_2'$,
6) $A \subseteq B' \Leftrightarrow B \subseteq A' \Leftrightarrow A \times B \in I$.

Повторное применение этого же оператора дает оператор $(\cdot)'' : 2^G \to 2^G$, удовлетворяющий следующим свойствам (для $X, Y \subseteq G$):

1) $X \subseteq Y \Rightarrow X'' \subseteq Y''$ (монотонность),
2) $X \subseteq X''$ (экстенсивность),
3) $(X'')'' = X''$ (идемпотентность).

Все перечисленные выше утверждения верны с точностью до замены множества объектов G на множество признаков M и наоборот, соответственно. Таким образом, оператор $(\cdot)''$ является оператором замыкания.

Пара (A, B) называется формальным понятием в формальном контексте (G, M, I), если $A \subseteq G$, $B \subseteq M$, $A' = B$, $B' = A$. Множество A называют объемом (extent) формального понятия, а множество B – содержанием (intent) формального понятия. Если представить формальный контекст в виде матрицы с булевыми значениями, то формальные понятия будут соответствовать максимальным подматрицам из единиц, которые можно получить из исходной с помощью перестановки строк и столбцов.

Множество всех формальных понятий формального контекста упорядочено отношением частичного порядка

$$(A_1, B_1) \geqslant (A_2, B_2) \Leftrightarrow A_2 \subseteq A_1 (\Leftrightarrow B_1 \subseteq B_2)$$

и образует полную решетку, называемую решеткой формальных понятий.

2.2. Многозначные диадические контексты

Многозначный диадический контекст принято описывать кортежем (G, M, W, I), где W является множеством значений контекста, т.е. множеством значений, которые признаки $m \in M$ могут принимать на объектах $g \in G$, а $I \subseteq G \times M \times W$. При этом если $(g, m, v) \in I$ и $(g, m, w) \in I$, то $v = w$. Такие контексты обычно представляются в виде таблиц, где в ячейке на пересечении строки, соответствующей объекту g, и столбца, соответствующего признаку m, записано значение $m(g) \in W$. В самом общем случае бикластером для таких данных является пара (A, B), где $A \subseteq G$, $B \subseteq M$ [2, 7].

2.3. Триадический случай и n-адический случай

Расширим приведенные выше определения на случай большей размерности. В этом разделе рассмотрим определения для общего случая большой размерности, а затем отдельно выведем представляющие интерес определения, необходимые для трикластеризации [8–11].

Во-первых, стоит обобщить понятие формального контекста. Пусть X_1, $X_2, \ldots, X_n$ — некоторые множества, а I — n-арное отношение, являющееся подмножеством их декартова произведения: $I \subseteq X_1 \times X_2 \times \ldots \times X_n$. В таком случае формальным n-адическим контекстом будет называться кортеж $K = (X_1, X_2, \ldots, X_n, I)$. В триадическом контексте $K = (G, M, B, I)$, по аналогии с диадическим случаем, множества G и M называются соответственно множествами объектов и признаков, а множество B называется множеством условий (от нем. Bedingungen — условия). Тогда тройка $(g, m, b) \in I$, где $g \in G$, $m \in M$, $b \in B$ может интерпретироваться как "объект g обладает признаком m при условии b".

Расширение оператора Галуа $(\cdot)'$ приобретает следующие разновидности: $2^{X_1} \to 2^{X_2} \times \ldots \times 2^{X_n}, \ldots, 2^{X_n} \to 2^{X_1} \times \ldots \times 2^{X_{n-1}}, \ldots, 2^{X_1} \times \ldots \times 2^{X_{n-1}} \to$ $\to 2^{X_n}, \ldots, 2^{X_2} \times \ldots \times 2^{X_n} \to 2^{X_1}$. Таким образом, в триадическом случае оператор производит переход из любого из трех множеств в декартово произведение оставшихся двух либо обратно.

Кортеж $(A_1, A_2, \ldots, A_n)$ называется n-адическим формальным понятием, если выполняется условие: $(A_1 \times \ldots \times A_{n-1})' = A_n, \ldots, (A_2 \times \ldots \times A_n)' = A_1$. В триадическом случае первые два элемента тройки множеств называются, как и в диадическом случае, объемом и содержанием соответственно. Третье множество называется модусом (от лат. Modus — мера, образ, способ). В n-адическом формальном контексте формальное понятие также является максимальным кубоидом.

В общем случае, в отличие от диадического, повторное применение оператора $(\cdot)'$ хотя и определено как $(\cdot)''$, не является оператором замыкания.

2.4. Многозначные триадические контексты

Дополним определение триадического контекста для многозначного случая. Пусть G, M, B — множества соответственно объектов, признаков и

условий. Пусть W — множество допустимых значений контекста. Тогда 4-арное отношение есть $I \subseteq G \times M \times B \times W$. Если кортеж $(g, m, b, w) \in I$, говорят, что "признак m принимает значение w на объекте g при условии b". $K = (G, M, B, W, I)$ называется многозначным формальным контекстом. Стоит заметить, что отношение I определено таким образом, что любому триплету $(g \in G,\ m \in M,\ b \in B)$ соответствует не более одного значения $w \in W$. Это можно представить в виде (частичной) функции означивания $V : Q \to W$, где $Q \subseteq G \times M \times B$. Тогда область определения Q этой функции будет называться областью определения многозначного формального контекста.

В общем случае трикластером называют тройку (X, Y, Z), где $X \subseteq G$, $Y \subseteq M$, $Z \subseteq B$ [3, 4] .

В методах, реализованных в рамках этой работы, рассматривается только случай, когда W является множеством вещественных чисел $\mathbb{R}$.

2.5. Критерии качества трикластеров

Для оценки качества обнаруживаемых трикластеров предлагается использовать показатели плотности и дисперсии.

Пусть $T = (X, Y, Z)$ — трикластер многозначного формального контекста $K = (G, M, B, W, I)$ с функцией означивания $V : Q \to W$. Определим плотность трикластера T как отношение количества входящих в него троек к его размеру:

$$\rho(T) = \frac{|Q \cap (X \times Y \times Z)|}{|X||Y||Z|}.$$

Можно заметить, что такая оценка не учитывает значения входящих в трикластер троек. Поэтому будем оценивать трикластеры еще и по дисперсии значений содержащихся в нем триплетов. Пусть $S = V(g, m, b) \mid (g, m, b) \in \in Q \cap (X \times Y \times Z)$ — выборка, состоящая из этих значений. Тогда несмещенная оценка выборочной дисперсии будет выглядеть так:

$$s^2 = \frac{\sum\limits_{i=1}^{|S|} S_i^2 - \left(\sum\limits_{i=1}^{|S|} S_i\right)^2 / |S|}{|S| - 1}.$$

Значения трикластера будем считать близкими при некотором параметре δ, если стандартное отклонение выборки будет меньше или равно параметру, т.е. $s \leqslant \delta$.

3. Обзор существующих методов многомерной кластеризации

Формальные понятия в бинарном контексте представляют собой максимальные кубоиды из единиц, что требует наличия достаточно жесткой струк-

туры. Это может быть полезно в контекстах без шума, пропущенных значений и ошибок, однако реальные данные редко соответствуют этим требованиям. Следовательно, при анализе в терминах формальных понятий с большой вероятностью будет происходить потеря некоторой части значимой информации, что приведет к получению нерелевантных результатов. Возможным решением этой проблемы является ослабление определения формального понятия, допускающее неполное заполнение структуры (теперь в нее также могут входить нули, "пропущенные тройки"). В общем случае такие сущности называются n-кластерами. В диадическом и триадическом случаях это соответственно би- и трикластеры.

Стоит заметить, что не существует единого определения трикластера, поэтому каждый раз он определяется исходя из потребностей конкретной задачи и порождающего метода [2, 3]. По этой причине в дальнейшем будем сравнивать эффективность различных порождающих методов с помощью таких общих параметров, как плотность, дисперсия и количество трикластеров [8].

3.1. OAC-трикластеризация

Задача OAC-трикластеризации довольно подробно описана в [4]: рассматриваются два метода, основанные на так называемых бокс- и штрих-операторах, и демонстрируется превосходство второго метода над первым по показателям плотности и разнообразия. По этой причине обратим внимание на последний, который, по сути, является расширением метода OA-бикластеризации, описанной в [7] на триадический случай. Для фиксированной тройки $(g, m, b) \in I$ выпишем для удобства штрих-операторы, используемые в методе, явно:

$$(g, m)' = \{b \mid (g, m, b) \in I\},$$
$$(g, b)' = \{m \mid (g, b, m) \in I\},$$
$$(m, b)' = \{g \mid (g, b, m) \in I\}.$$

Тогда OAC-трикластером, основанным на штрих-операторах, построенным для тройки $(g, m, b) \in I$, будет называться тройка множеств $T = (((m, b)', (g, b)', (g, m)')$.

Отличительной особенностью построенных таким образом трикластеров является наличие в них плотной подструктуры, $(m, b)' \times \{m\} \times \{b\}, \{g\} \times \times (g, b)' \times \{b\}, \{g\} \times \{m\} \times (g, m)'$, трехмерного креста из "единиц" [4].

Алгоритм построения списка трикластеров описывается следующим образом: для каждой тройки $(g, m, b) \in I$ пополняются двумерные массивы, содержащие результаты выполнения операций штрих, например, $Prime[g, m] = = Prime[g, m] \cup b$, сам трикластер содержит три указателя на соответствующие массивы $T = (^*Prime[m, b],^* Prime[g, b],^* Prime[g, m])$.

Сложность алгоритма по времени (в худшем случае) составляет $O(|I|)$. Однако самой ресурсоемкой процедурой становится проверка уникальности трикластера, требующая, в худшем случае, сравнения нового трикластера

со всеми уже найденными. Этот шаг можно ускорить за счет использования различных способов хеширования и эффективного хранения информации для сравнения.

3.2. Шкалирование формальных понятий и метод TriMax

Метод шкалирования формальных понятий (Conceptual Scaling), рассмотренный в [5], предполагает поиск бикластеров близких значений в многозначных контекстах с помощью средств анализа формальных понятий (АФП).

Пусть $K = (G, M, W, I)$ — многозначный диадический формальный контекст. Определим бикластер близких значений с параметром θ как бикластер, где $\forall g_i, g_l \in G$, $\forall m_i, m_k \in M \;\; |m_i(g_j) - m_k(g_l)| \leqslant \theta$, т.е. все значения принадлежат одному классу толерантности ($m_i(g_j) \simeq_\theta m_k(g_l)$).

Далее в [4] предлагается выделить из множества W классы толерантности, сопоставить каждому из них формальный контекст, содержащий только элементы, входящие в данный класс, и произвести в них поиск формальных понятий, также являющихся бикластерами близких значений. Эта идея обрела воплощение в виде алгоритма TriMax [5] .

В приложении к проблеме, решаемой в данной работе, главным недостатком этого метода является использование инструментария анализа формальных понятий, действующего только в диадическом случае.

3.3. Межпорядковое шкалирование

Метод, рассмотренный в подразделе 3.2, зависит от параметра θ и ищет только бикластеры, в которых значения отстоят друг от друга не более чем на величину параметра. В этом подразделе познакомимся с методом межпорядкового шкалирования (Interordinal Scaling), предложенным в [5] и позволяющим производить поиск бикластеров близких значений по всем доступным значениям параметра сразу с помощью инструментов триадического анализа формальных понятий (Triadic Concept Analysis — TCA). Этот подход заключается в дополнении многозначного диадического формального контекста до шкалированного триадического. Третьим измерением в таком случае становятся интервалы, в которые входит значение, принимаемое признаком на объекте.

Пусть (G, M, W, I) — многозначный диадический формальный контекст. Построим новое интервальное измерение (scale dimension) T из W с помощью процедуры межпорядкового шкалирования (Interordinal Scaling). Шкала представляет собой бинарное отношение $J \subseteq W \times T$, сопоставляющее каждому исходному значению из W соответствующие элементы из T. Пусть $T = \{[\min(W), w], \forall w \in W\} \cup \{[w, \max(W)], \forall w \in W\}$. Тогда $(w, t) \in J$ тогда и только тогда, когда $w \in t$, где $t \subseteq T$.

Пусть $Y \subseteq G \times M \times T$ — тернарное отношение. В таком случае $(g, m, t) \in$ $\in Y$ тогда и только тогда, когда $m(g) \in t$. Кортеж (G, M, T, Y) называется

шкалированным триадическим контекстом многозначного диадического контекста (G, M, W, I).

В [5] доказывается, что кортеж (A, B, U), где $A \subseteq G$, $B \subseteq M$, $U \subseteq T$, является триадическим формальным понятием, только если (A, B) является бикластером близких значений для некоторого $\theta \geqslant 0$.

Этот метод довольно сложно адаптировать для поиска трикластеров близких значений в многозначных контекстах, так как он требует поиска формальных понятий в контекстах размерности большей, чем исходная, а соответствующие инструменты для тернарного случая исследованы лишь в общем виде.

4. Предложенные методы

Как было продемонстрировано в предыдущей части, существующие методы многомерной кластеризации не приспособлены решать поставленную в данной работе задачу, но содержат идеи, опираясь на которые можно спроектировать требуемый метод. В этом разделе предлагаются два алгоритма для поиска трикластеров близких значений в многозначных триадических контекстах. Первый является расширением OAC-трикластеризации на многозначный случай, а второй представляет собой разновидность классического алгоритма одномерной кластеризации k-средних, в котором используется предложенная первым автором данной работы метрика.

4.1. Метод NOAC

Метод NOAC (Numerical OAC) получен модификацией OAC-трикластеризации, основанной на штрих-операторах. Он принимает параметр δ, который определяет, какие значения считаются близкими. Пусть даны многозначный триадический контекст $K = (G, M, B, W, I)$ и (частичная) функция означивания $V : G \times M \times B \to W$, переводящая триплет (g, m, b) в соответствующее значение w тогда и только тогда, когда $(g, m, b, w) \in I$. Область определения этой функции обозначим как $Q \subseteq G \times M \times B$. Как и в бинарном случае, будем строить трикластер от фиксированной тройки $(\tilde{g}, \tilde{m}, \tilde{b}) \in Q$. Переопределим операторы, используемые методом:

$$(\tilde{g}, \tilde{m})^\delta = \left\{ b \;\middle|\; (\tilde{g}, \tilde{m}, b) \in Q \wedge |V(\tilde{g}, \tilde{m}, b) - V(\tilde{g}, \tilde{m}, \tilde{b})| < \delta \right\},$$
$$(\tilde{g}, \tilde{b})^\delta = \left\{ m \;\middle|\; (\tilde{g}, m, \tilde{b}) \in Q \wedge |V(\tilde{g}, m, \tilde{b}) - V(\tilde{g}, \tilde{m}, \tilde{b})| < \delta \right\},$$
$$(\tilde{m}, \tilde{b})^\delta = \left\{ g \;\middle|\; (g, \tilde{m}, \tilde{b}) \in Q \wedge |V(g, \tilde{m}, \tilde{b}) - V(\tilde{g}, \tilde{m}, \tilde{b})| < \delta \right\}.$$

Назовем эти операторы δ-операторами. Тогда OAC-трикластером, основанным на δ-операторах, построенном на тройке $(\tilde{g}, \tilde{m}, \tilde{b}) \in Q$, будет называться тройка множеств $T = \left((\tilde{m}, \tilde{b})^\delta, (\tilde{g}, \tilde{b})^\delta, (\tilde{g}, \tilde{m})^\delta \right)$.

Можно заметить, что построенный таким образом трикластер содержит плотную трехмерную подструктуру из близких значений по аналогии с немногозначным случаем. В силу того что результат применения δ-операторов зависит от конкретной тройки, в алгоритме NOAC не используется предподсчет. Для каждой тройки $(g, m, b) \in Q$ совершим два шага:

1) Вычислим множества $(g, m)^\delta$, $(g, b)^\delta$ и $(m, b)^\delta$.
2) Если трикластер $T = ((m, b)^\delta, (g, b)^\delta, (g, m)^\delta)$ еще не содержится в множестве трикластеров, добавим его туда.

Перебор всех троек в худшем случае займет $O(|I|)$. Построение трикластера производится за $O(|G| + |M| + |B|)$. Следовательно, сложность алгоритма можно оценить как $O(|I| \cdot \max(|G|, |M|, |B|))$. Отсутствие предподсчета освобождает от необходимости хранить лишние данные, поэтому сложность по памяти составит $O(|I|)$, что необходимо для хранения всех троек. Проверка уникальности вычисленного трикластера существенно усложняется с ростом общего количества трикластеров (в худшем случае $O(|I| \log(|I|))$.

4.2. Метод Tri-k-means

Для сравнения предложенного выше метода с альтернативным вариантом был выбран классический алгоритм одномерной кластеризации k-средних с авторской метрикой, учитывающей близость по компонентам входного триконтекста. Пусть $K = (G, M, B, W, I)$ — многозначный триадический контекст. Пусть функция $V : G \times M \times B \to W$ переводит тройку (g, m, b) в соответствующее значение w тогда и только тогда, когда $(g, m, b, w) \in I$, а $Q \subseteq$ $\subseteq G \times M \times B$ – ее область определения.

Расстояние между тройками $t_1 = (g_1, m_1, b_1) \in Q$ и $t_2 = (g_2, m_2, b_2) \in Q$ вычисляется по формуле

$$\rho(t_1, t_2) = |V(t_1) - V(t_2)| + \gamma([g_1 \neq g_2] + [m_1 \neq m_2] + [b_1 \neq b_2]),$$

где γ является вещественным параметром, определяющим приоритетность близости значений внутри трикластера относительно расширения его по измерениям. Выражение $[a_1 \neq a_2]$ принимает значение 0, если координаты тройки в соответствующем измерении эквивалентны, и 1 иначе.

В результате работы алгоритма получаем k кластеров, состоящих из троек. Для сравнения с методом NOAC требуется перевести их в трикластеры. Пусть $H \subseteq Q$ – множество триплетов, входящих в кластер. Тогда трикластером близких значений будем называть тройку множеств

$$T = (\{g \mid \exists (g, m, b) \in H\}, \{m \mid \exists (g, m, b) \in H\}, \{b \mid \exists (g, m, b) \in H\}).$$

Таким образом, в соответствующие измерения получившегося трикластера войдут все значения, содержащиеся в тройках исходного кластера. Эту адаптацию классического алгоритма k-средних было решено назвать методом Tri-k-means. Вычислительная сложность одной итерации алгоритма k-средних линейно зависит от параметра k и общего количества трикластеров, по-

этому может быть оценена как $O(k \cdot |G||M||B|)$. Пусть i – количество итераций алгоритма, тогда время работы всего алгоритма в худшем случае можно оценить как $O(k \cdot i \cdot |G||M||B|)$.

5. Эксперименты

Исходный код всех инструментов, подробное описание данных и результаты экспериментов доступны по ссылке: `https://github.com/EgurnovD/TriclusteringToolbox`.

5.1. Описание данных

Утилитой ContextGenerator был создан эталонный контекст, состоящий из 1120 троек, сформированных в два кубоида со значениями 3 и 7. Для него построены наборы контекстов с пропущенными значениями (10–90%) и размытием значений с различной амплитудой (0,1–2).

Контекст с реальными данными был создан на основе набора данных 100k проекта GroupLens [12] (`https://grouplens.org/datasets/movielens/100k/`), содержащего информацию о 100 000 оценках по пятибалльной шкале, поставленных 1000 пользователями сайта MovieLens 1700 фильмам при наличии 19 жанров.

5.2. Результаты экспериментов

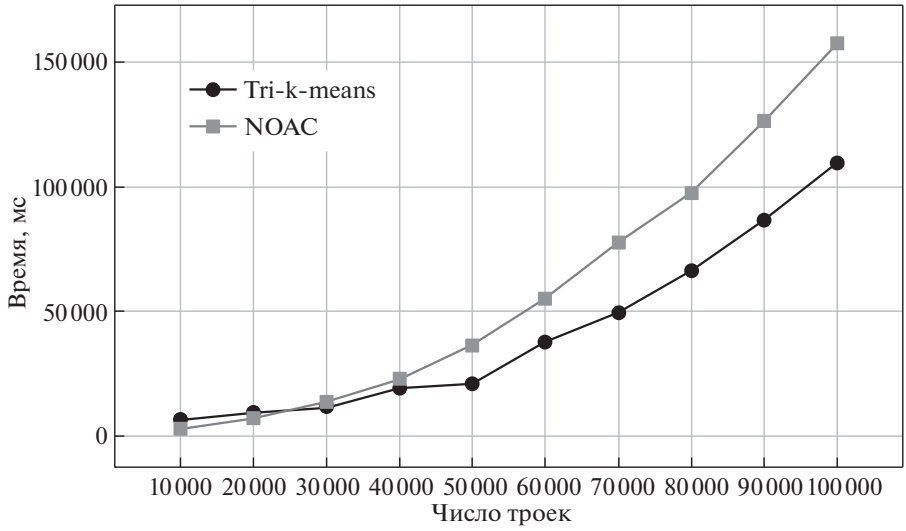
В первой серии экспериментов проверялась устойчивость алгоритмов к отсутствию данных. Метод NOAC смог найти эталонные трикластеры без дополнительной обработки при уровне потерь 10%. Метод Tri-k-means справился даже с 90% отсутствием данных за счет предварительного знания о количестве трикластеров (параметр k).

Вторая серия экспериментов была посвящена проверке на шумоустойчивость, в которой метод NOAC выдавал трикластеры с достаточно близкими значениями до тех пор, пока диапазоны размытия не пересеклись. Трикластеры, вычисленные методом Tri-k-means, в среднем имели большую дисперсию и перестали удовлетворять условию близости значений при меньшей амплитуде размытия.

На реальных данных и метод NOAC, и метод Tri-k-means генерируют трикластеры, удовлетворяющие условию близости значений, но метод NOAC порождает значительно более плотные трикластеры. Их можно интерпретировать как клубы по интересам, состоящие из пользователей, примерно одинаково оценивших схожие по жанрам фильмы. Отсутствующие значения в таком случае можно применить в рекомендательной системе, предполагая, что новые оценки будут дополнять существующие трикластеры, не сильно отклоняясь от среднего значения.

6. Заключение

В данной работе были рассмотрены современные методы би- и трикластеризации, а именно OAC-трикластеризация, шкалирование формальных



Зависимость времени работы от объема выборки.

понятий (Conceptual Scaling) и межпорядковое шкалирование (Interordinal Scaling), и предложено два алгоритма для поиска трикластеров близких значений в многозначных триадических контекстах. Один из них, являющийся расширением OAC-трикластеризации на многозначный случай, назван методом NOAC. Другой представляет собой классический алгоритм кластеризации k-средних и использует авторскую метрику для вычисления расстояний (метод Tri-k-means). Эксперименты показали, что на синтетических контекстах метод NOAC лучше справляется с размытием значений. Из реальных данных он извлекает более качественные трикластеры в терминах плотности, но тратит на это немного больше времени (см. рисунок). Стоит отметить, что метод NOAC принимает только один легко интерпретируемый параметр.

СПИСОК ЛИТЕРАТУРЫ

1. *Zaki M.J., Meira Jr W.* Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Cambridge University Press, 2020.
2. *Madeira S.C., Oliveira A.L.* Biclustering algorithms for biological data analysis: a survey // IEEE/ACM Trans. on Comp. Biol. and Bioinf., IEEE. 2004. V. 1. No. 1. P. 24–45.
3. *Henriques R., Madeira S.C.* Triclustering algorithms for three-dimensional data analysis: a comprehensive survey // ACM Computing Surveys (CSUR), ACM. 2018. V. 51. No. 5. P. 1–43.
4. *Ignatov D.I., Gnatyshak D.V., Kuznetsov S.O., Mirkin B.G.* Triadic formal concept analysis and triclustering: searching for optimal patterns // Machine Learning, Springer. 2015. V. 101. No. 1. P. 271–302.
5. *Kaytoue M., Kuznetsov S.O., Macko J., Napoli A* Biclustering meets triadic concept analysis // Ann. Math. Artif. Intell., Springer. 2014. V. 70. No. 1. P. 55–79.
6. *Ganter B., Wille R.* Formal Concept Analysis: Mathematical Foundations. Berlin/ Heidelberg: Springer, 1999.

7. *Ignatov D.I., Kuznetsov S.O., Poelmans J.* Concept-based biclustering for internet advertisement // In proc. ICDMW, IEEE. 2012. P. 123–130.
8. *Lehmann F., Wille R.* A triadic approach to formal concept analysis // Conceptual Structures, LNCS, Springer. 1995. V. 954. P. 32–43.
9. *Voutsadakis G.* Polyadic Concept Analysis // Order, Springer Verlag. 2002. V. 19. No. 3. P. 295–304.
10. *Ganter B., Obiedkov S.* Implications in triadic formal contexts // In proc. International Conference on Conceptual Structures, Springer. 2004. P. 186–195.
11. *Ananias K.H.A., Missaoui R., Ruas P.H.B., Zarate L.E., Song M.A.J.* Triadic concept approximation // Information Sciences, Elsevier. 2021. V. 572. P. 126–146.
12. *Harper F.M., Konstan J.A.* The movielens datasets: History and context // ACM Transact. on Interactive Intell. Syst., ACM. 2015. V. 5. No. 4. P. 1–19.

Статья представлена к публикации членом редколлегии О.П. Кузнецовым.

Поступила в редакцию 20.01.2021

После доработки 07.02.2022

Принята к публикации 15.02.2022