Нейский И.М., Филиппович А.Ю.

Методика адаптивной кластеризации фактографических данных на основе интеграции алгоритмов MST и Fuzzy C-means

Актуальной задачей развития современных предприятий является переход к «экономике, основанной на знаниях»[1]. Для ее решения необходимо определиться с тем, какие знания есть в предприятии, как ими пользоваться, как они влияют на бизнес-результаты и как ими эффективно управлять.

Исследования показывают, что значительная часть знаний содержится не только в памяти сотрудников, но и в рамках информационного поля компании, которое представлено многочисленными документами, данными корпоративных информационных систем, ресурсами Интернет и другими источниками. Постоянно растущий объем информации, неструктурированный и неявный характер представленных таким образом знаний требует разработки специализированных методик и программных инструментов.

[1]Статья подготовлена при поддержке гранта Президента РФ № МК-5341.2007.9

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.

Подтверждением этого служит то, что на международной конференции "Ситуационные центры. Методы. Решения. Реализация", прошедшей в Российской академии государственной службы при Президенте РФ весной 2008 г., вопросы использования средств Data Mining для построения ситуационных моделей стали одной из самых обсуждаемых тем. Среди наиболее проблематичных недостатков подобных решений были отмечены:

- разнородность моделей, используемых в различных информационно-аналитических системах и отсутствие механизмов их интеграции;

- низкий уровень адаптируемости типовых методик для решения задач в различных предметных областях и динамически изменяемых ситуациях;
- в алгоритмах систем недостаточно проработаны вопросы учета НЕ-факторов (неопределенности, нечеткости, неверности и т.д.)

Предложенная в статье Методика адаптивной кластеризации фактографических данных на основе интеграции алгоритмов MST и Fuzzy C-means направлена на устранение указанных недостатков для одной из самых распространенных задач интеллектуального анализа данных (Data mining).

Потребность в кластеризации возникает в тех областях/этапах деятельности, где есть необходимость в разбиении объектов (ситуаций) на непересекающиеся подмножества, называемыми кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Четкое разделение на кластеры возможно только в идеальных условиях и при сильно различающихся параметрах объектов кластеризации. Поэтому для решения реальных задач все чаще применяются нечеткие методы, в которых разбиение объектов (ситуаций) выполняется на частично пересекающиеся подмножества.

Важной предпосылкой применения нечетких методик кластеризации в реальных условиях является то, что характеристики объектов не всегда являются измеримыми и поэтому в ряде случаев присутствуют экспертные оценки характеристик объектов, которые являются субъективными и могут быть противоречивыми.

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.

Исследование существующих методов кластеризации выявило следующие недостатки:

- Алгоритмы работают преимущественно с одним типом, как правило, с числовыми данными;
- Большинству алгоритмов характерна определенная форма выделяемых кластеров;
- Точное количество выделяемых кластеров является необходимым и обязательным параметром пуска алгоритма;
- Чувствительность к выбросам и отклонениям;
- Медленная работа на больших объемах данных;
- Линейность/нелинейность в большую сторону времени работы алгоритма в зависимости от объема входных данных;

- Большая вычислительная сложность;
- Возникновение неопределенностей при обработке и распределении объектов по кластерам.

Анализ большого количества методов кластеризации подтверждает гипотезу о том, что методы кластерного анализа являются контекстно-зависимыми и имеют превосходство в решении определенного круга задач. Применение того или иного метода также определяется линейной делимостью набора данных или наличия нелинейных взаимосвязей. 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.

Можно выделить следующие факторы, влияющие на выбор метода кластеризации:

- необходимая точность выполнения кластеризации;
- априорное представление о количестве кластеров;
- типы данных исследуемых атрибутов;
- объем анализируемой информации;
- время выполнения анализа;
- оценочная оптимальность полученного результата;
- равномерность исследуемой информации;
- разделяемость исходных данных.

В зависимости от характеристик исходных данных и желаемого результата используют следующие математические аппараты:

- статистика;
- деревья решений;
- нейронные сети;
- теория графов;
- нечеткая логика.

Статистические методы, как и деревья решений, теория графов, применяются в задачах, когда данные хорошо разделимы и не содержат большого количества выбросов. В других случаях предпочтение отдают нейронным сетям и нечеткой логике. Методы, основанные на нечеткой логике, позволяют делать мягкое разделение массива данных на кластеры, определяя одни и те же данные в разные кластеры с разной степенью принадлежности. Эта особенность позволяет применять методы на неоднородных данных и получать при этом довольно качественные результаты.

Нейронные сети и нечеткая логика являются лидерами среди методов анализа данных большого объема со значительным количеством анализируемых атрибутов.

Для выбора наилучшего метода кластеризации разработаны следующие критерии:

- Критерий 1: Объем информации (количество строчек или по занимаемому месту) по отношению к времени обработки.
- Критерий 2: Размерность информации (количество атрибутов в строке) в порядковом выражении.
- Критерий 3: Типы атрибутов: числовой дискретный, числовой непрерывный, строковый.
- Критерий 4: Чувствительность к равномерности информации (наличие аномалий-выбросов во входном наборе данных).
- Критерий 5: Априорное (экспертное) представление о форме получаемых кластеров.
- Критерий 6: Априорное (экспертное) представление о количестве кластеров.
- Критерий 7: Априорное (экспертное) представление о перекрываемости кластеров.
- Критерий 8: Качество кластеризации.

Приведенные критерии и сравнение методов по этим критериям подтверждают гипотезу контекстной зависимости методов от данных и сложность задачи выбора «идеального» метода для того или иного набора данных. Также остается нерешенной в общем случае задача наличия в наборах данных различных типов данных, в т.ч. лингвистического типа. Результат кластеризации можно улучшить с помощью применения нескольких алгоритмов кластеризации и предварительных подготовительных процедур, связанных с очисткой и нормализацией входного набора данных. Сложность кластеризации заключается не только в выделении кластеров, но и в последующей оценке полученных результатов. Также на выбор метода кластеризации влияет наличие практической реализации того или иного метода в виде программного модуля.

Задача выбора метода кластеризации должна начинаться с выбора используемого инструмента, определяющего качество, скорость и конечность результата кластеризации, а затем уже останавливаться на самом методе кластеризации, который уже внесет коррективы в получаемый результат.

Как правило, при выполнении кластеризации необходимо настраивать методы кластеризации для исследуемой предметной области в зависимости от исходных данных. Кустомизирующие параметры можно разделить на две группы:

- Характеристические;
- Итерационные;

- Экспертные.

Характеристические параметры используются для общей оценки входного набора, например, количество записей, количество атрибутов данных, тип атрибутов данных, используемость атрибутов в проведении исследования и др.

Итерационные параметры характеризуются тем, что точное значение параметра заведомо неизвестно и подбирается итерационным перебором в выделенном интервале значений.

Экспертные параметры используются для более точной настройки алгоритмов, в состав данного вида параметров входят такие параметры, как количество кластеров, коэффициент отталкивания и др. Величину параметра можно получать эмпирическим путем или итерационным, оценивая результаты кластеризации. Данный вид параметров, как и итерационные параметры, требуют от аналитика наличия определенного опыта и знания специфики предметной области в довольно значительном объеме.

При исследовании девяти алгоритмов (CURE, BIRCH, MST, k-средних, PAM, CLOPE, Самоорганизующиеся карты Кохонена, HCM, Fuzzy C-means) были выявлены следующие параметры:

Характеристические параметры: объем обучающего множества, объем валидационного множества, объем тестового множества, количество, тип, используемость атрибутов входного набора данных.

Итерационные параметры: количество кластеров, алгоритм выполнения дополнительной кластеризации, пороговое значение остановки работы алгоритма, способ выбора начальных центров, максимальное количество итераций, количество одновременно обрабатываемых данных, количество предварительных разделов, коэффициент удаленности.

Экспертные параметры: способ определения расстояния между кластерами, метод оценки качества кластеризации, пороговое значение для метода оценки качества кластеризации, начальное пороговое значение алгоритма, процент аномалий (выбросов) в полном объеме, разделяющая функция, скорость обучения сети.

Большое число параметров требуют наличия значительного опыта аналитика в предметной области и знания специфики исходных данных, используемых в аналитическом исследовании. Использование значений «по умолчанию» может привести к низкому качеству кластеризации и получению неадекватных результатов даже при правильном выборе метода кластеризации. Отсутствие и недостаточность опыта в предметной области можно компенсировать оценкой проводимой кластеризации.

Использование аналитических алгоритмов кластеризации позволит снизить риск неправильного принятия решения и человеческого фактора ошибки, поскольку сам алгоритм принятия решения будет зафиксирован в виде алгоритма, а экспертные оценки параметров запуска алгоритма будут усредняться и уточняться в процессе применения алгоритма.

На основании приведенных выше фактов и высказываний методика адаптивной кластеризации должна удовлетворять следующим требованиям:

- обеспечивать приемлемую точность выполнения выделения кластеров;
- не требовать точного указания кластеров;
- работать с основными типами данных: числовыми, лингвистическими;
- выделять кластеры произвольной формы;
- обладать пониженной чувствительностью к выбросам и отклонениям в выборке данных;
- обладать приемлемым временем работы с большими объемами данных и нелинейностью в меньшую сторону при увеличении объемов в большую сторону (масштабируемость);
- отсутствие неопределенностей при распределении объектов по кластерам;
- обладать средней степенью вычислительной сложности;
- иметь настраиваемые параметры пуска для адаптации методики к особенностям предметной области.

После анализа существующих методик и алгоритмов для решения поставленной задачи из инструментов выполнения кластеризации были выбраны: теория графов и нечеткая логика.

Определяющими в выбранной комбинации были способность при использовании графов выделять кластеры произвольной формы и оптимальной структуры, при использовании математического аппарата нечеткой логики выполнить разделение объектов с лингвистическими атрибутами. При разработке новой методики адаптивной кластеризации за основу взяты принципы, используемые в алгоритмах MST [1] и Fuzzy C-means [2].

Двухэтапность выполнения кластеризации и использование оценочной функции разбиения позволяет повысить качество проводимой кластеризации. Вычисление глобального критерия делает алгоритм кластеризации во много раз быстрее, чем при использовании локального критерия при парном сравнении объектов, поэтому «глобализация» оценочной функции – один из путей получения масштабируемых алгоритмов.

Еще одним достоинством нечеткой кластеризации является то, что использование нечеткости при определении объектов по кластерам позволит сделать более полное разбиение исходного множества на кластеры, ликвидируя тем самым неопределенности, которые возникают при четком разбиении.

В связи с выявленными требованиями, особенностями и критериями предложен следующий алгоритм «ADAKL»:

Входные данные алгоритма:

$D = \{u_1, u_2, ..., u_m\}$, 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 $j^{го}$ атрибута $i^{го}$ 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 видов значений атрибутов: $ValueTypes = \{Целочисленный\ тип, Денежный\ тип, Лингвистический\ тип\}$, где $Целочисленный\ тип \subset \mathbb{Z}$, $Денежный\ тип \subset R$, $Лингвистический\ тип \subset Словарная\ система$;

$Словарная\ система = \{Лингв.тип1, Лингв.тип2, ..., Лингв.типS\}$, где $Лингв.тип_i$ – объект словарной системы, характеризующий оценочные/качественные показатели объектов кластеризации;

Множество типов значений атрибутов: $FieldTypes = \{Входное, Идентифицирующее, Информационное\}$, где «*Входное*» – означает участие атрибута объекта в алгоритме, «*Идентифицирующее*» – обозначает ключевой атрибут объектов кластеризации, идентифицирующий каждый объект входного набора данных, «*Информационное*» – обозначает атрибут объекта, не оказывающий влияние на результаты работы алгоритма.

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:

$$Metrics = \left\{ \begin{matrix} Евклидово\ расстояние, Квадрат\ Евклидова \\ расстояния, расстояние\ Чебышева \end{matrix} \right\};$$

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

$$OstTreeMethods = \left\{ \begin{matrix} Алгоритм\ Борувки, Алгоритм \\ Крускала,\ \ Алгоритм\ Прима \end{matrix} \right\};$$

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

$$NormMethods = \{Линейная\ нормализация, Статистическая\ нормализация\}$$

Выходные данные алгоритма:

$$C = \left\{ C_1, C_2, ..., C_c \middle| O^c \to max, c \le q, C_1 \cup C_2 \cup ... \cup C_c = D \right\},$$

$u_i \overset{\mu_{ij} \in [0,1]}{\in} C_j$, $i = \overline{1,m}$, $j = \overline{1,c}$

Описание алгоритма:

Начало.

Этап 1. Нормализация значений числовых атрибутов.

В случае линейной нормализации выполняется следующее:

$$Value_{ij} := \left\{ \frac{Value_{ij}}{\underset{j}{Max}\left(Value_{ij}\right)} \middle| \begin{matrix} \underset{j}{Max}\left(Value_{ij}\right) \ne 0, \\ t_j \in \{Целочисленный\ тип, Денежный\ тип\} \end{matrix} \right\}$$

В случае статистической нормализации выполняется следующее: $Value_{ij} := \left\{ \frac{Value_{ij} - \frac{\sum_{i=1}^{m} Value_{ij}}{m}}{\sqrt{\frac{\sum_{i=1}^{m} \left(Value_{ij}\right)^2}{m} - \left(\frac{\sum_{i=1}^{m} Value_{ij}}{m} \right)^2}} \middle| \begin{matrix} \underset{j}{Max}\left(Value_{ij}\right) \ne 0, \\ t_j \in \left\{ \begin{matrix} Целочислен- \\ ный\ тип, Де- \\ нежный\ тип \end{matrix} \right\} \end{matrix} \right\}$

[4]

Этап 2. Вычисление матрицы взаимных расстояний между объектами.

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

Если $Metric = Евклидово\ расстояние$, то

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

Если $Metric = Квадрат\ Евклидова\ расстояния$, то

$$Dist_{ij} = \sum_w \left(\left[Value_{iw} - Value_{jw} \right] * K_w \right)^2,$$

Если $Metric = Расстояние Чебышева$, 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 $FieldType[w] = "Входное"$

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 $i^{го}$ объекта к $j^{му}$ 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: $Dist_{ij}^{k} := \left\{ 0 \middle| Dist_{ij}^{k} \overset{i,j=\overline{1,m}}{=} Max \right\}$.

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:

$$Avg[r] = \frac{\sum_{u_j \in V_i^k} \left\{Value_{jr} \middle| FieldType[w] = "Входное"\right\}}{\left|V_i^k\right|}, \qquad j = \overline{1,m},$$

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

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:

$$Avg[r] = \left\{ \underset{\sum \left\| Value_{jr} - Value_{lr} \right\| = min \text{ или } \underset{u_j \in V_i^k, u_l \in V_i^k}{\varphi = max}}{Value_{jr}} \middle| FieldType[w] = "Входное" \right\},$$

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

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

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

Если $Metric = Расстояние\,Чебышева$, 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 $FieldType[w] = "Входное"$.

Шаг 5. Нормализация матрицы расстояний от объектов до центров кластеров V_i^k: $Dist_{ij}^{k'} = \begin{cases} \dfrac{Dist_{ij}^k}{Max\left(Dist_{ij}^k\right)}, Max\left(Dist_{ij}^k\right) \neq 0 \\ 1, Max\left(Dist_{ij}^k\right) = 0 \end{cases}$, $i = \overline{1,k}$, $j = \overline{1,m}$.

Шаг 6. Соотнесение объектов к кластерам в соответствии со степенью удаленности элементов кластера (w): $u_j \in V_i^k \left| Dist_{ij}^{k'} \le w \text{ или } Dist_{ij}^{k'} = \underset{i}{Min}\left(Dist_{ij}^{k'}\right)\right.$, $i = \overline{1,k}$, $j = \overline{1,m}$.

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:

$\mu_{ij} = \dfrac{\mu_{ij}}{\sum_{i=1}^{k} \mu_{ij}}$, $j = \overline{1,m}$.

Шаг 9. Вычисление центров полученных кластеров с использованием матрицы нечеткого разбиения [2]:

$$V_i^{k'} = \frac{\sum_{j=1}^{m} \mu_{ij}^{\ p} * u_j}{\sum_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} \mu_{ij}^{\ p}}, i = \overline{1,k}.$$

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

Шаг 10. Оценка качества полученного разбиения.

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

$$O^k = \frac{\sum_{i=1,k} \frac{\left|V_i^{k'}\right| * \sum_{j=1}^{m} \mu^p{}_{ij} * \left\|V_i^{k'} - u_j\right\|_{u_j \in V_i^{k'}}}{\underset{i \neq j}{Min}\left(\left\|V_i^{k'} - u_j\right\|_{u_j \in V_i^{k'}}\right) * Max\left(\left\|V_i^{k'} - u_j\right\|_{u_j \in V_i^{k'}}\right) * \sum_{j=1}^{m} \left\|V_i^{k'} - u_j\right\|_{u_j \in V_i^{k'}} * k}}{m * k^2}$$, 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: $O_{Onm} = \underset{i=1,q}{MAX}\left(O^i\right)$.

Конец.

Предложенный алгоритм 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 алгоритма 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

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

- использует критерий оценки разбиения на кластеры с учетом специфики предметной области: небольшое количество кластеров, наибольшая плотность, средняя удаленность объектов.

Описанный алгоритм имеет практическую ценность в применении данной методики для выделения групп клиентов брокерского обслуживания профессионального участника на финансовых и фондовых рынках.

В процессе опытно-промышленной эксплуатации алгоритма было проведен анализ работоспособности алгоритма в следующих исследованиях:

- Исследование 1: выделение секторов инвестирования на основе анализа показателей финансовых инструментов;
- Исследование 2: выделение групп клиентов на основе данных об оборотах за период, частоты проведения операций, количестве финансовых инструментов, группы используемых финансовых инструментов;
- Исследование 3: выявление категорий финансовых инструментов для оценки эффективности операций;
- Исследование 4: выделение классов автомобилей на основе данных о максимальной скорости, цвете кузова, воздушном сопротивлении, массе.

Показатель / Метод		Количество полученных кластеров	Оценка разбиения[2]	Оценка разбиения с заданным количеством кластеров (без учета лингвистических атрибутов)	Оценка разбиения с заданным количеством кластеров (с учетом лингвистических атрибутов)
1	Метод 1	30	1.0000	-	1.0000
	Метод 2	-	-	-	-
	Метод 3	30	1.0000	-	1.0000
2	Метод 1	12	0.1667	0.7467	0.7467
	Метод 2	-	-	0.6000	-
	Метод 3	3	1.0000	1.0000	1.0000
3	Метод 1	15	1.0000	0.8953	1.0000

[2] Оценка разбиения выполняется на основе показателей выполненной кластеризации: $O = \frac{r}{n} \times \begin{cases} q/k, q \le k \\ k/q, q > k \end{cases}$, где

q – количество кластеров по итогам кластеризации;

r – количество элементов, правильно распределенных по соответствующим кластерам;

k – исходное количество кластеров;

n – количество объектов кластеризации.

	Метод 2	-	-	0.6388	-
	Метод 3	10	1.0000	0.9857	1.0000
4	Метод 1	9	0.1111	1.0000	0.7778
	Метод 2	-	-	1.0000	-
	Метод 3	3	1.0000	1.0000	1.0000

где Метод 1 – карта Кохонена, Метод 2 – алгоритм к – средних, Метод 3 – ADAKL.

Средние оценки проведенных экспериментов:

Оценка / Метод	Оценка разбиения	Оценка разбиения с заданным количеством кластеров (без учета лингвистических атрибутов)	Оценка разбиения с заданным количеством кластеров (с учетом лингвистических атрибутов)	Итоговая оценка
Метод 1	0.5695	0.8807	0.8811	0.7771
Метод 2	-	0.7463	-	0.2488
Метод 3	1.0000	0.9952	1.0000	0.9984

Алгоритм ADAKL показал наилучший результат в проведенных исследованиях. Три исследования проводились на основе данных деятельности кредитной организации, а одно исследование проводилось на данных о технических характеристиках автомобилей. Предложенный алгоритм показал лучший результат во всех проведенных исследованиях, что подтверждает возможность использования его в решении поставленных задач кредитной организации, а также его адаптивность к использованию в других предметных областях.

Особое внимание при адаптации алгоритма к работе в той или иной предметной области необходимо обратить на следующие входные параметры алгоритма: весовой коэффициент влияния атрибута объекта, размазанность кластеров, степень удаленности элементов, способ определения расстояния между объектами, способ проведения нормализации значений числовых атрибутов.

Весовой коэффициент позволяет выделять отдельную статистическую информация, которая играет наибольшую роль при выделении групп объектов. Размазанность кластеров позволяет делать более четкими или наоборот более размазанными границы между группами объектов. Также значительную роль при выделении групп объектов играет способ расчета информационного расстояния между объектами, а впоследствии и кластерами, с помощью которого можно сделать анализ «гладким», применив Евклидовы метрики, или сделать анализ «грубым», применив метрики с математическими функциями минимизации и максимизации.

Способ нормализации является дополнительным механизмом улучшения качества кластеризации, который особенно актуален при наличии в данных каких – либо выбросов и отклонений. Сте-

пень удаленности элементов является регулятором идентичности элементов при определении объектов по группам, влияющим на количество итоговых групп.

Разработанный алгоритм адаптивной кластеризации позволяет проводить оперативный анализ клиентской базы с целью оптимизации затратных статей компании и повышения эффективности проведения операций на фондовом рынке. На базе алгоритма разработан программный комплекс. В процессе адаптации используемых математических методов к предметной области разработан специализированный метод оценки расстояний, в алгоритм внедрен приоритетный метод анализа параметров исходного набора, позволяющий использовать данный алгоритм для разработки тарифной политики обслуживания клиентов. Разработанная методика кластеризации позволяет выполнять кластеризацию исходных данных по лингвистическим характеристикам объектов.

В данном алгоритме используется параметр «Степень удаленности элементов», который позволяет ограничивать степень свободы алгоритма при определении объектов в кластеры и оказывать влияние на количество выделяемых кластеров. Параметр «Размазанность кластеров» позволяет сглаживать или уточнять границы получаемых кластеров.

Выполнение кластеризации адаптированным для этой предметной области методом позволит расширить перечень имеющихся групп клиентов, отойти от текущего порядка оценки операций/групп клиентов и расширить клиентскую базу, что, в дальнейшем, позволит повысить доходность выделенного направления банковской деятельности и разнообразить перечень предоставляемых услуг клиентам.

Данная методика также применима для других предметных областей, в которых имеется множество объектов с заданными характеристиками. Использование комплексного подхода повышает качество результатов кластеризации, повышает вероятность получения конечного решения поставленной задачи и ускоряет процесс выполнения кластеризации, а также упрощает процесс взаимодействия система – пользователь за счет предварительной обработки входных данных. Данная работа найдет применение в следующих исследованиях: анализ клиентской базы (кредиты физических лиц), ранжирование существующих финансовых инструментов, проведение анализа отраслей с целью инвестирования ресурсов и др.

Литература

1. Huahai He, Ambuj K. Singh « Efficient Algorithms for Mining Significant Substructures in Graphs with Quality Guarantees ». Источник: Seventh IEEE International Conference On Data Mining.

2. С.Д. Штовба «Введение в теорию нечетких множеств и нечеткую логику». Источник: http://matlab.exponenta.ru/fuzzylogic/book1/index.php.

3. Jeff Erickson. Minimum spanning trees. Lecture notes: Fall 2002 — CS 373: Combinatorial Algorithms. Источник: http://compgeom.cs.uiuc.edu/~jeffe/teaching/373/notes/13-mst.pdf.

4. А.А. Ежов, С.А. Шумский «Нейрокомпьютинг и его применения в экономике и бизнесе». – Курс лекций: Лекция 7. Источник: http://www.intuit.ru/department/expert/neurocomputing/7/neurocomputing_7.html