УДК 681.518:658.18

В.В. Сайко

КЛАСТЕРИЗАЦИЯ ДАННЫХ МАРКЕТИНГОВЫХ ИССЛЕДОВАНИЙ

Введение

В настоящее время широкое распространение получила технология Data Mining - технология добычи знаний из большого массива данных. Одним из главных математических инструментов, используемых этой технологией, является кластерный анализ входного множества данных. При этом кластер рассматривается как набор объектов, имеющих существенное сходство между собой и отличающихся от объектов, принадлежащих другим кластерам.

Кластерный анализ все чаще применяется в маркетинговых исследованиях. При этом решается, в первую очередь, проблема группировки. Речь может идти о продуктах, клиентах, рынках сбыта. Однако, программные средства, наиболее полно реализующие этот аппарат (SPSS, STATISTICA), [1, 2] требуют как высокой математической подготовки исследователя, так и определенного опыта в аналитической работе на конкретном рынке. Каждый раз перед аналитиком стоит не только задача выбора адекватного алгоритма кластеризации, но и проблема квалифицированной интерпретации результатов.

Чаще всего применение кластерного анализа в маркетинговых исследованиях связано с задачей сегментации. Решение этой задачи приводит к идентификации устойчивых групп потребителей, каждая из которых объединяет в себя индивидуумов с похожими характеристиками.

Исследователи отмечают, что формальный подход к этой процедуре приводит к ошибкам в подборе характеристик, на основе которых проводится кластеризация. Главным критерием часто является доступность данных, а не их информативность. Однако включение даже небольшого количества незначимых характеристик может привести к неадекватному разбиению на кластеры [3].

Особый интерес, в приложении к маркетинговым исследованиям, вызывают алгоритмы кластеризации для качественных шкал. К качественным шкалам относят:

а) номинальную шкалу (обычно используется для регистрации пола, рода занятий, обычного места покупки того или иного товара и т.д.)

б) порядковую шкалу (экспертные оценки, оценки предпочтений, шкала возрастных интервалов и др.)

На сегодняшний день предложено свыше десятка методов для работы с качественными данными [4]: например, семейство иерархических кластерных алгоритмов. Одним из наиболее эффективных считается алгоритм LargeItem, который основан на оптимизации некоторого глобального критерия. В общем случае наличие глобального критерия дает возможность гораздо быстрее проводить кластерный анализ, чем при использовании локального критерия при парном сравнении объектов, поэтому "глобализация" оценочной функции считается наиболее перспективным путем получения масштабируемых алгоритмов.

На этом же принципе основан алгоритм CLOPE, предложенный в 2002 году группой китайских ученых. Он обеспечивает более высокую производительность и лучшее качество кластеризации в сравнении с алгоритмом LargeItem и многими иерархическими алгоритмами [5, 6].

В основе алгоритма CLOPE лежит идея максимизации глобальной функции стоимости, которая повышает близость транзакций в кластерах при помощи увеличения параметра кластерной гистограммы.

С помощью параметра, названного авторами CLOPE коэффициентом отталкивания, регулируется уровень сходства транзакций внутри кластера, а, значит, и окончательное количество кластеров. Этот коэффициент подбирается пользователем. Чем больше коэффициент отталкивания, тем ниже уровень сходства и тем больше кластеров будет сгенерировано.

Таким образом, анализ существующих алгоритмов кластеризации позволил выявить ряд проблем, актуальных для такой предметной области, как автоматизация маркетинговых исследований:

 ISSN 1562-9945

При использовании существующих программных средств возможны ошибки в подборе характеристик (в частности, включение в анализ малозначимых переменных), на основе которых проводится кластеризация, что приводит к неадекватному разбиению на кластеры.

Большинство алгоритмов требует начального указания количества кластеров, что не всегда оправдано с практической точки зрения и может привести к потере более удачного разбиения.

Выбор алгоритмов и метрик требует высокой квалификации исследователя.

Постановка задачи

Задача настоящего исследования состоит в том, чтобы разработать эффективный алгоритм для решения задач кластеризации данных маркетинговых исследований, измеренных на качественных шкалах. Решение этой задачи является частью проекта создания специализированного программного обеспечения для проектирования маркетинговых исследований, а также ввода и анализа полученных данных. Исходные данные далее полагаются структурированными и классифицированными в соответствии с типами шкал и видами вопросов.

Результаты

Рассмотрим основные элементы предлагаемой методики.

Формирование и редукция признакового пространства

В разрабатываемой системе принята пятиуровневая классификация вопросов анкеты, изображенная на рис. 1 (окружностями отмечены наблюдаемые переменные Dk, представленные набором значений {Pk1,…,Pki} (частоты выбора респондентами i-того варианта ответа), i=2…10. Переменные Dk измерены на качественных шкалах. Уровни организованы таким образом, что содержат статистически независимые переменные. Базовая переменная Dб выбирается среди вопросов первого уровня – в ее терминах формируется описание будущего кластера. Значения базовой переменной далее именуются категориями респондентов или категориями базовой переменной. В данной работе рассматривается пример с базовой переменной «Возраст респондентов», принимающей 10 значений: «до 18 лет»,…, «более 65 лет».

Общий принцип формирования набора признаков показан на рис. 1. По указанной схеме последовательно осуществляется кросс-табуляция базовой и прочих переменных исследования Dk. Пропущенные данные заменяются средними значениями соседних категорий. Переменные ранжируются в соответствии с условием:

$$\mathrm{cov}(D_б, D_k) \to \max \qquad (1)$$

Далее, в соответствии с (1) выделяются 4 переменных, наиболее информативных с точки зрения последующей кластеризации. При этом выделенные переменные Dk могут быть измерены на порядковых или номинальных шкалах, их табулированные значения Pkij, для каждой j-той категории базовой шкалы имеют интервальный характер (см. рис. 2).

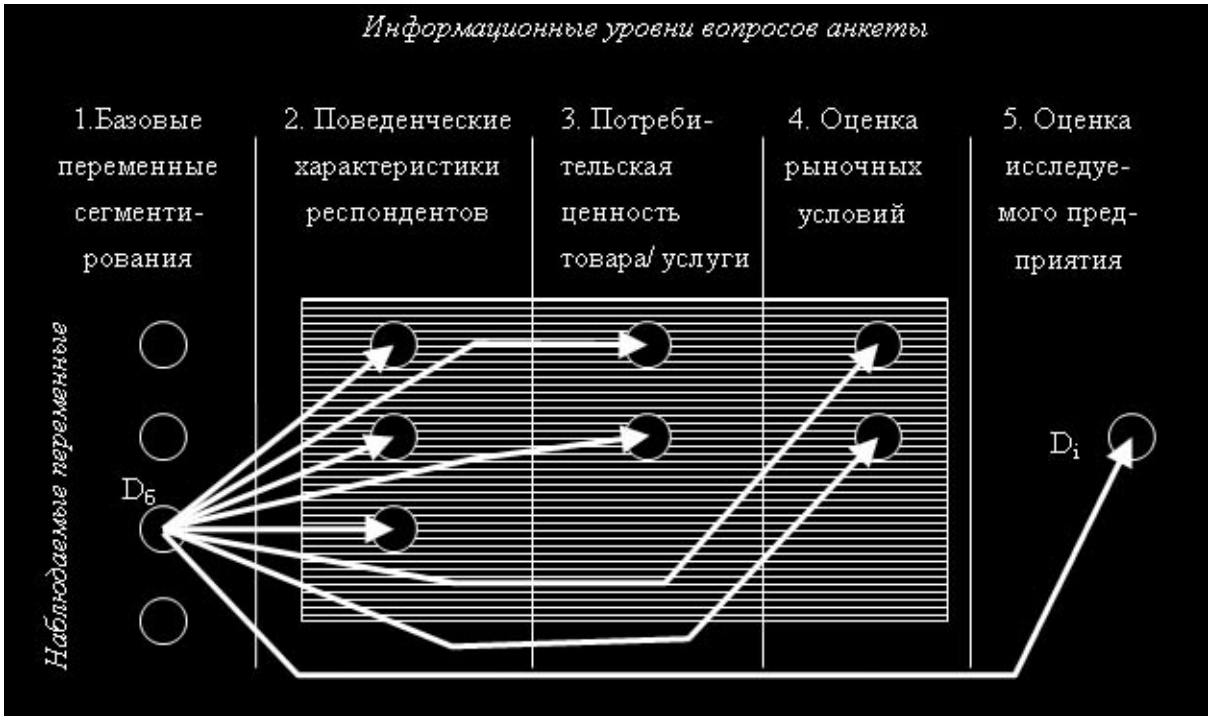

Рисунок 1 - Уровневая организация данных маркетингового исследования и организация поиска значимых данных

С целью исключения шума, в каждой выбранной переменной исключаются из рассмотрения значения, для которых:

$$Д_{ij} = \min(P_{ij}^{\max}) - \max(P_{ij}^{\min}) \diamondsuit 0 \qquad (2)$$

здесь $P_{ij}^{\max}, P_{ij}^{\min}$ - верхние и нижние значения интервала соответственно.

Следует отметить, что обычно с использованием (2) исключаются из анализа частоты выбора респондентами таких вариантов ответов, как «затрудняюсь ответить», «нет».

 ISSN 1562-9945

Далее значения переменной D_k ранжируются в соответствии с критерием:

$$k_k = \frac{Д_{kij}}{\min(P_{kij}^{\max})} \to \max . \qquad (3)$$

Таким образом, коэффициент k_k отражает разделяющую силу значения P_{kij} и для каждой из четырех выявленных ранее переменных можно получить набор модифицированных значений k_kP_{kij}, отражающих характер отклика каждой j-той категории базовой переменной на i-тый вариант ответа k-того вопроса анкеты (рис. 3).

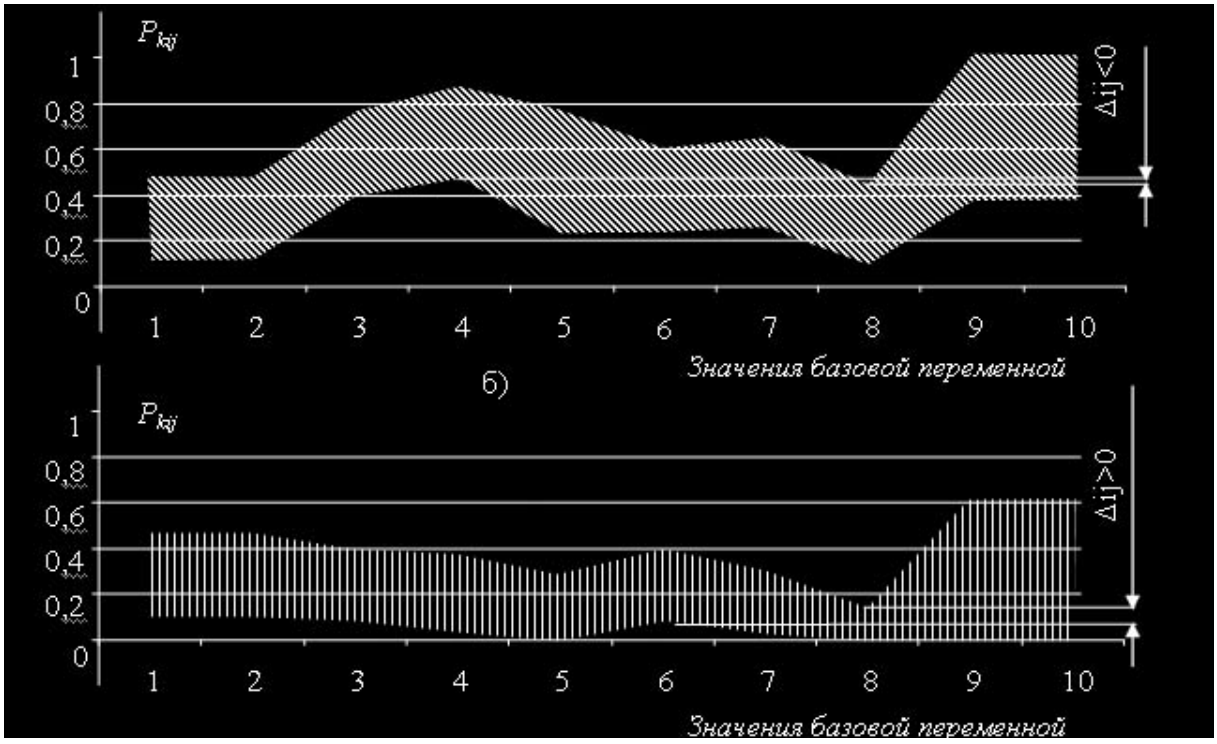

Рисунок 2 - Примеры включаемых в анализ (а) и исключаемых из анализа (б) значений переменных, полученных в результате кросс-табуляции

Таким образом, итоговое значение P_k переменной D_k можно представить в виде произведения векторов, скалярные значения которых равны k_kP_{kij}, исходящих из одной точки под углами, кратными π/m, где m – количество оставшихся после редукции значений переменной D_k. Тогда общую характеристику категорий базовой переменной можно представить лепестковой диаграммой, приведенной на рис. 4. Последующая свертка количества измерений путем получения для каждой категории значений P_1–P_2 и P_3–P_4 позволяет представить исходные данные в виде, приведенном на рис.5.

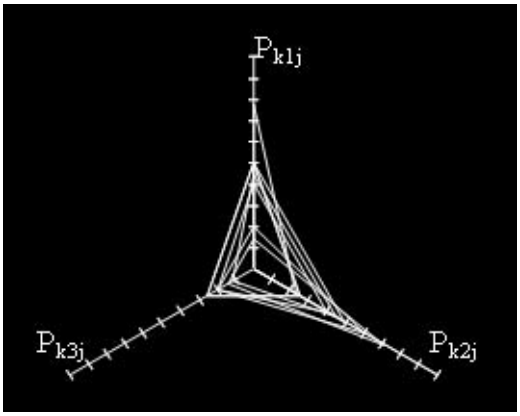

Рисунок 3 - Распределение модифицированных значений переменных для десяти категорий базовой шкалы

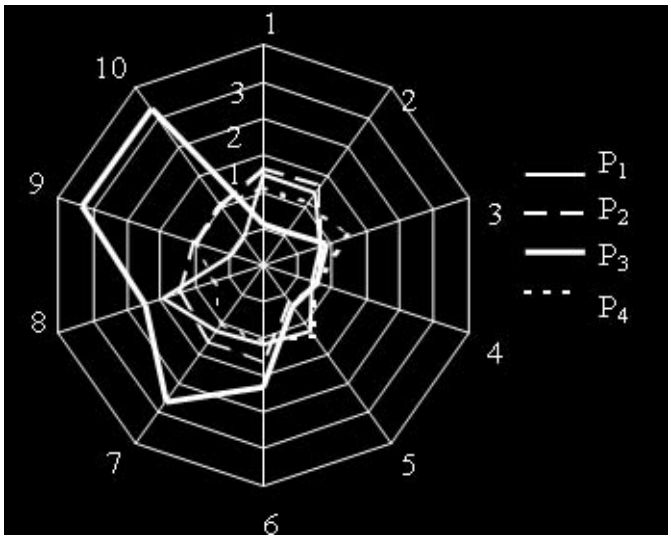

Рисунок 4 - Общая характеристика 10 категорий базовой переменной с использованием модифицированных переменных

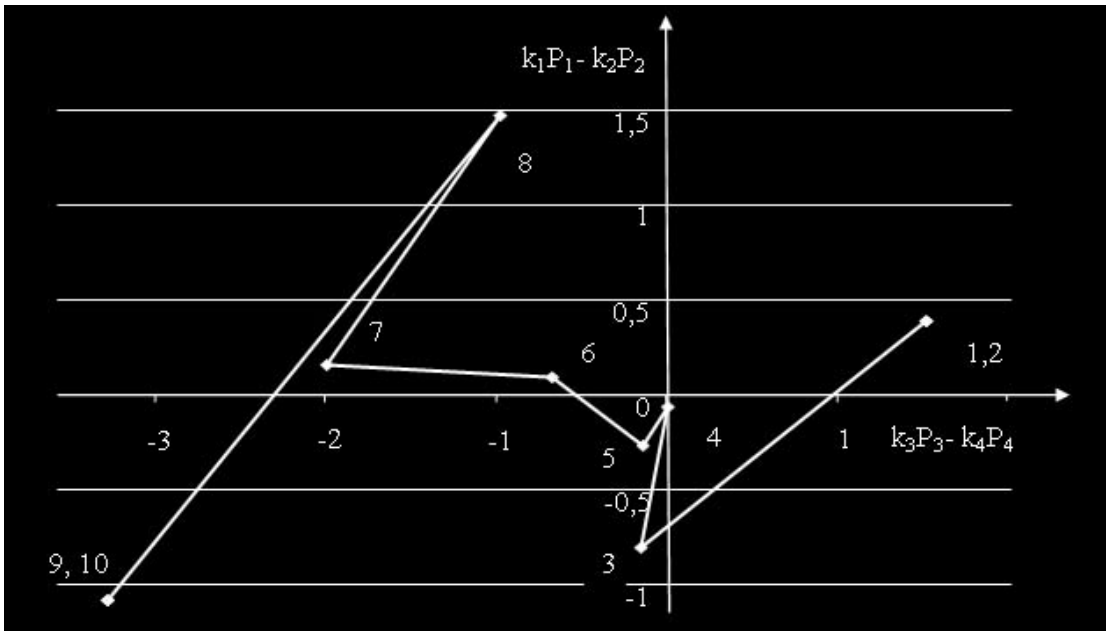

Рисунок 5 - Результат редукции признакового пространства (точки соединены в порядке следования категорий базовой переменной)

Математическая постановка задачи

Таким образом, постановку задачи кластеризации с использованием порядковой шкалы можно сформулировать следующим образом: имеется множество векторов $I=\{I_1,I_2,\ldots,I_n\}$, где n – количество категорий базовой переменной. Каждый из векторов характеризуется углом αn, отражающим характер реакции респондентов, отнесших себя к n-той категории, на вопросы анкеты 2-5 уровней (свойство n-той категории респондентов) и длиной ln

 ISSN 1562-9945

(степень выраженности свойства). Необходимо осуществить разбиение указанного множества на m кластеров. При этом количество кластеров не задано, однако практическую ценность имеют разбиения с $1 < m \le 4$. Каждый сформированный кластер G_m может состоять из одного и более элементов. Кластеры в данном случае представляют собой секторы, ограниченные углами $\alpha_m^{\min}$ и $б_m^{\max}$. Пример исходных данных с основными обозначениями приведен на рис. 6.

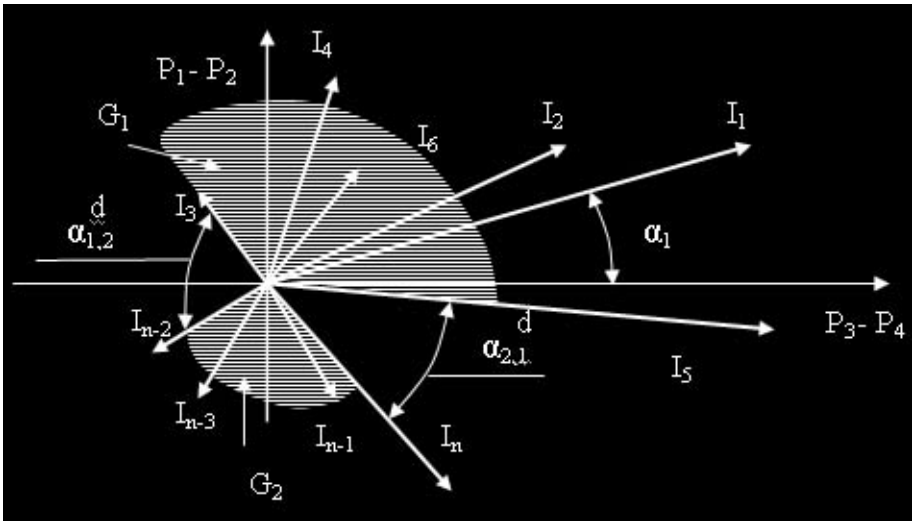

Рисунок 6 - Математическая постановка задачи кластеризации множества векторов I

Условимся, что

$$б_m^{\max} - б_m^{\min} < \frac{p}{2}, \qquad (4)$$

исключив, таким образом, возможность включения в одну группу категорий респондентов с полярно противоположным откликом на вопросы анкеты.

Из определения порядковой шкалы следует, что крайние векторы набора I не могут принадлежать одному и тому же кластеру:

$$\exists G_1(I_1 \exists G_1 \exists I_n \exists G_m), \qquad (5)$$

$$\exists G_m(I_1 \exists G_1 \exists I_n \exists G_m). \qquad (6)$$

Тогда справедливо следующее утверждение:

$$\exists(I_k, I_{k+1} | k > 1 \exists k+1 < n \exists I_k \exists G_{m\exists 1} \exists I_{k+1} \exists G_m) \exists |б_k - б_{k+1}| = б_{m-1,m}^d, \qquad (7)$$

где $\alpha_{m-1,m}^d$ - величина, отражающая степень различия кластеров G_{m-1} и G_m (разделяющий угол). Из выражений (5, 6) также следует:

$$б_{m-1,m}^d \exists б_{1,n}. \qquad (8)$$

Понятие схожести векторов в данной постановке задачи целесообразно дополнить условиями:

$$б^{d}_{m\text{-}1,m} \leftarrow \frac{б^{\max}_{m} - б^{\min}_{m}}{\mathrm{card}G_{m} - 1} \leftarrow \max, \quad (9)$$

$$б^{d}_{m\text{-}1,m} > \min(б_{1,2}, б_{2,3}, ..., б_{n\text{-}1,n}). \quad (10)$$

Предложенная форма представления данных позволяет сформулировать еще одну важную закономерность, действующую для порядковой шкалы:

$$\leftarrow(I_k \left| \leftarrow(I_i, I_{i+1} \right| (б_i \leftarrow б_k \leftarrow б_{i+1}))) \leftarrow \{I_i, I_{i+1}, ..., I_k\} \leftarrow G'_m, \quad (11)$$

где G'_m - субкластер, границы которого пока не определены. Данное правило выполняется, например для вектора I_6 и пары I_2, I_3 на рис. 6, что позволяет сделать вывод о безусловной принадлежности $\{I_2,..., I_6\}$ к одному кластеру.

Из (11) следует также правило конкатенации субкластеров:

$$\leftarrow(G', G'' \left| \leftarrow I_i \right| (I_i \leftarrow G' \leftarrow I_i \leftarrow G'')) \leftarrow \{G', G''\} \leftarrow G'_m. \quad (12)$$

Результаты алгоритмизации

Алгоритм кластеризации, основанный на (4-12) можно представить в виде одновременного распространения очагов инфекции $\{I_i,...,I_k\}$, выявленных с помощью правила (11) в наборе векторов I. При этом оценивается иммунитет всех векторов, примыкающих к очагам, и выбирается один из них, обеспечивающий требование, вытекающее из выражения (9):

$$б^{\max}_{m} - б^{\min}_{m} \leftarrow \min \quad (13)$$

Безусловно, присоединяется к кластеру вектор, для которого

$$б^{\min}_{m} \leftarrow б_i \leftarrow б^{\max}_{m}. \quad (14)$$

Вектор считается устойчивым к инфекции, т.е. определенно не относящимся к данному кластеру, если выполняется одно из выражений:

$$б_i - б^{\min}_{m} \leftarrow \frac{p}{2}, \text{ если } б_i \leftarrow б^{\max}_{m}, \quad (15)$$

$$б^{\max}_{m} - б_i \leftarrow \frac{p}{2}, \text{ если } б_i \leftarrow б^{\min}_{m}. \quad (16)$$

Рассмотрим четыре возможных варианта остановки развития очагов инфекции:

1. Одна из границ кластера поглощает первый или последний вектор набора I, а вторая граница не может расшириться в соответствии с условиями (15, 16). В этом случае кластер считается полностью сформированным (см. рис. 7,а).

2. Обе границы кластера совмещаются с первым и последним вектором набора I. В этом случае кластеризация считается невыполнимой.

3. Отсутствуют векторы, не поглощенные каким-либо кластером. Кластеризация считается завершенной.

4. Очередной шаг развития инфекции может привести к слиянию двух субкластеров. Это возможно лишь при выполнении условия (4). В противном случае границы соседних очагов считаются определенными (см. рис. 7,б).

Результаты работы предлагаемого алгоритма на конкретной задаче представлены на рис. 8. В данном случае, на основе анализа переменных «Распределение покупательской активности», «Частота приобретения одежды», «Факторы, влияющие на выбор одежды», «Факт приобретения одежды в магазине N» удалось выделить две возрастных категории респондентов – «до 40 лет» и «от 40 лет и старше», существенно отличающиеся друг от друга с точки зрения потребительского поведения. При формировании маркетинговой политики следует учитывать, что первая категория (наиболее активные покупатели) в меньшей степени интересуются исследуемым магазином, чем вторая, гораздо более пассивная аудитория.

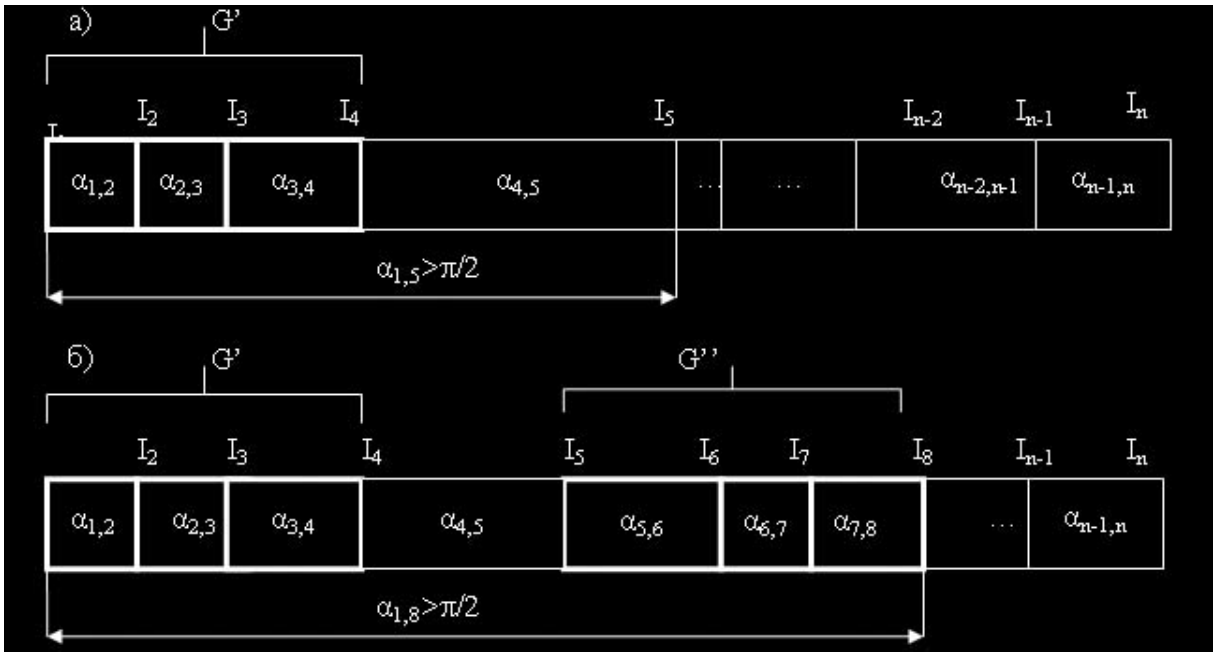

Рисунок 7 - Варианты завершения процесса развития субкластеров

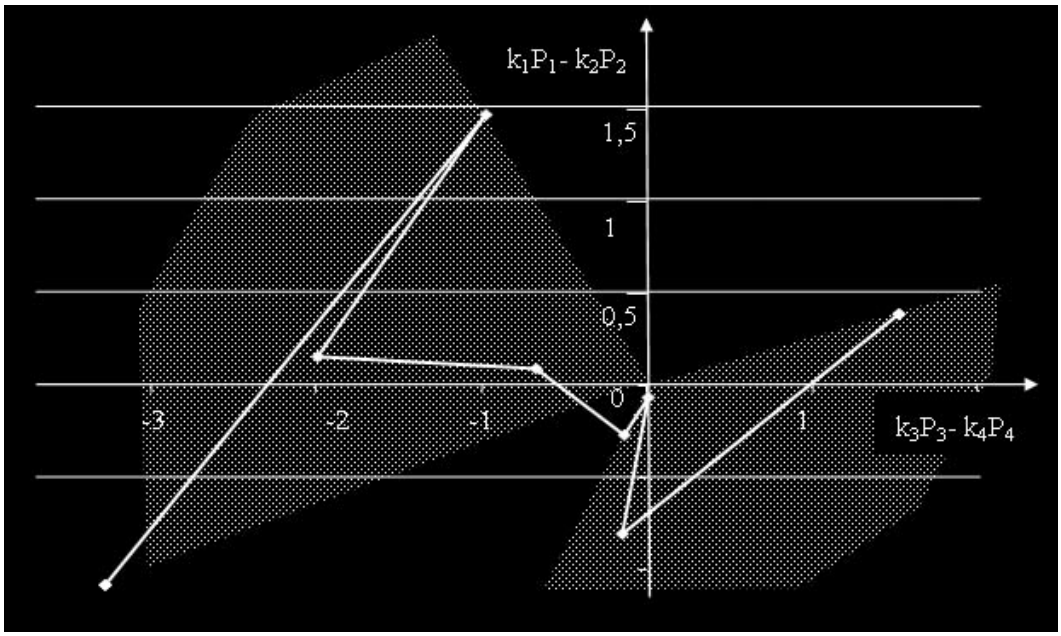

Рисунок 8 - Пример результата кластеризации

Программная реализация

В разрабатываемом программном обеспечении результаты кластеризации интерпретируются как новая латентная переменная исследования. Пользователю предлагается уточнить ее наименование и наименования ее значений (кластеров) – см. рис. 9. Полученная переменная может использоваться в ходе дальнейшего анализа данных.

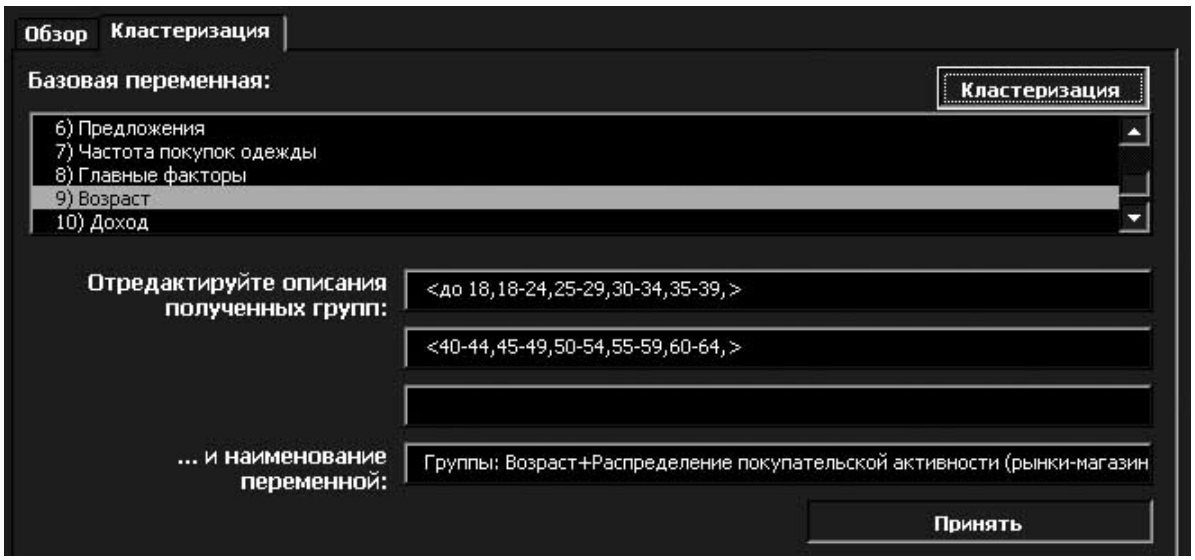

Рисунок 9 - Этап корректировки наименований групп, полученных в результате кластеризации

 ISSN 1562-9945

Выводы

В данной работе, на основе анализа предметной области удалось сформулировать постановку задачи кластеризации данных, полученных в ходе маркетинговых исследований. Предложена эффективная методика формирования и редукции признакового пространства кластеризации. Сформулированы и обоснованы правила кластеризации данных, измеренных на порядковых шкалах. Разработан алгоритм кластеризации, основанный на идее развития очагов инфекции – фрагментов признакового пространства, безусловно принадлежащих к одному и тому же кластеру. Алгоритм не требует предварительного указания количества кластеров, определяет его самостоятельно и выполняет разбиение за один проход. Программная реализация алгоритма в рамках разработанной автором системы проектирования маркетинговых исследований и анализа полученных данных подтверждает его работоспособность и практическую ценность. В настоящее время продолжается работа над процедурой кластеризации с использованием номинальных шкал.

ЛИТЕРАТУРА

1. Кузнецов Д.Ю., Трошина Т.Л. Кластерный анализ и его применение // http://www.yspu.yar.ru/vestnik/uchenue_praktikam/33_4/
2. Рыбалко В. В. Параметрическое диагностирование энергетических объектов на основе факторного анализа в среде Statistica // Exponenta Pro.— 2004. — N 2.— с. 78-83.
3. Ларин С. В. Выявление обобщенных ассоциативных правил // Exponenta Pro.— 2003.— N 3.— с. 34-38.
4. Punj Girish, Stewart David W. Clustering Algorithms In Marketing Research // Journal of Marketing Research, Vol. XX, (May 1983), pp.134-148
5. Мандель И.Д. Кластерный анализ / И.Д. Мандель. - М.: Финансы и статистика, 1988. - 176 с.
6. Кластеризация категорийных данных: масштабируемый алгоритм CLOPE // http://www.basegroup.ru/ tasks/datamining_prepare.htm
7. Yang, Y., Guan, H., You. J. CLOPE: A fast and Effective Clustering Algorithm for Transactional Data In Proc. of SIGKDD’02, July 23-26, 2002, Edmonton, Alberta, Canada.
8. Wang, K., Xu, C.. Liu, B. Clustering transactions using large items. In Proc. CIKM’99, Kansas, Missouri, 1999.

Получено 21.01.2008 г.