Кластерный анализ детализаций телефонных переговоров

Б.И. Рабинович
Институт проблем информатики РАН

Аннотация

В работе представлен один из видов многомерного статистического анализа - кластерный анализ. Проведено исследование, в котором рассматривается применение различных сочетаний метрик и алгоритмов кластерного анализа к детализациям телефонных переговоров - одного из источников данных Логико-Аналитической системы «Аналитик». Актуальность задачи заключается в том, что результаты исследования могут быть использованы в решении таких важных задач, как: выявлении преступных групп лиц в криминалистике, разработке новых тарифов и предложений в мобильной связи, выявлении групп счетов в банковском секторе и т.п.

Annotation

The article is devoted to the one of types of multidimensional statistical analysis – cluster analysis. In this research an application of the different combination of metrics and algorithms of cluster analysis to billings of telephone calls is considered. Billings is one of all sources of data in Logical-oriented Analytic System “Analyst”. The urgency of this task is in the results of the research, which could be used in the solution of such important tasks as showing up criminal elements in criminalistics, development of new pricing plans and proposals in mobile communication, revealing groups of accounts on banking area and so on.

Введение

Одной из актуальных задач аналитической обработки структурированной информации является обработка детализаций телефонных переговоров. В рамках Логико-Аналитической Системы «Аналитик» [29], разработанной на базе ИПИ РАН, был предложен анализ «Детализация номерных объектов», состоящий из нескольких аналитических режимов обработки детализаций телефонных переговоров и банковских переводов [35]. Для расширения возможностей системы предлагается рассмотреть применение к детализациям один из видов многомерного статистического анализа – кластерного анализа [15]. С помощью этого типа анализа решаются такие важные задачи, как классификация и группировка объектов по множеству их атрибутов, что в свою очередь позволит выявить, например, связанные между собой телефонные номера или банковские счета в детализациях. Реализация кластерного анализа в системе «Аналитик» позволит существенно сократить время обработки массивов биллинговой информации аналитиками вручную, автоматизировав этот процесс.

В статье даны основные определения и описаны сущности рассматриваемой предметной области, проведено исследование существующих метрик и алгоритмов кластеризации, рассчитаны результаты их использования в различных комбинациях при кластеризации детализаций. На основании полученных результатов сделаны выводы об эффективности использования различных комбинаций метрик и алгоритмов кластеризации.

1 Основные определения предметной области.

1.1 Кластерный анализ.

Кластерный анализ - совокупность математических методов, предназначенных для формирования относительно "отдаленных" друг от друга групп "близких" между собой объектов по информации о расстояниях или связях (мерах близости) между ними, и по смыслу аналогичен терминам «автоматическая классификация», «таксономия», «распознавание образов без учителя» [27]. Фактически "кластерный анализ" - это обобщенное название достаточно большого набора алгоритмов, используемых при создании классификации. В ряде изданий используются такие синонимы кластерного анализа, как классификация и разбиение. Кластерный анализ широко используется как средство типологического анализа. В любой научной деятельности классификация является одной из фундаментальных составляющих, без которой невозможны построение и проверка научных гипотез и теорий [41].

Значимость кластерного анализа в задачах анализа данных описана в работах [4; 6; 10; 13]. Техника кластеризации применяется в самых разнообразных областях. Хартиган Д. А. [9] дал обзор многих опубликованных исследований, содержащих результаты, полученные методами

кластерного анализа. Существуют примеры применения кластеризации в медицине, археологии, маркетинге, биологии и в задачах распознавания графических образов [2; 14; 31]. Таким образом, когда необходимо классифицировать большие объемы информации на пригодные для дальнейшей обработки группы, кластерный анализ оказывается полезным и эффективным [42].

В том или ином объеме методы кластерного анализа имеются в большинстве наиболее известных отечественных и зарубежных статистических пакетах: Sigamd, DataScope, Stadia, СОМИ, ПНП-БИМ, СОРРА-2, СИТО, SAS, SPSS, Statistica, BMDP, Statgraphics, Genstat, S-Plus и т.д. Достаточно подробный сравнительный анализ многочисленный статистических пакетов можно найти в [25]. Данный обзор был сделан в 80 гг., с того времени появились новые версии многих статистических программ, появились и абсолютно новые программы, использующие как новые алгоритмы, так и сильно возросшие мощности вычислительной техники. Однако большинство статистических пакетов используют алгоритмы, предложенные и разработанные в 60-70 гг. В работах [18; 39; 38, стр. 88-92] описано более 60 статистических программных продуктов, в том числе наиболее распространенные в настоящий момент программы статистического анализа, такие как: Statistica [19; 36], Stadia [30], Statgraphics [22], Stata [http://www.stata.com].

1.2 Основные обозначения и определения.

Пусть множество $I=\{I_1, I_2, \ldots, I_n\}$ состоит из n объектов (индивидов):, принадлежащих, некоторой популяции π_I. Предположим также, что существует некоторое множество *наблюдаемых* показателей или атрибутов $C= (C_1, C_2, \ldots, Cz)^T$, которыми обладает каждый индивид из I. Наблюдаемые атрибуты могут быть как *количественными,* так и *качественными*; однако основная часть исследования будет посвящена количественным данным, которые иногда будем называть *измерениями.* Результат измерения a-й характеристики I_b объекта будем обозначать через x_{ab}, а вектор $X_b=[x_{1b}\times x_{2b}\times..\times x_{zb}]$ размерности $z\times 1$ будет отвечать каждому ряду измерений (для b-го индивида). Таким образом, для множества индивидов I исследователь располагает множеством векторов измерений $X=\{X_1, X_2, \ldots, X_n\}$, которые описывают множество I. Отметим, что множество X может быть представлено как n точек в z-мерном евклидовом пространстве E_z [24, стр. 5-7].

Пусть m — целое число, меньшее, чем n. Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся в множестве X, разбить множество объектов I на m кластеров (подмножеств) $\pi_1, \pi_2, \ldots, \pi_m$ так, чтобы каждый объект I_b принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время как объекты, принадлежащие разным кластерам, были разнородными.

Решением задачи кластерного анализа является разбиение, удовлетворяющее некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок. Этот функционал часто называют целевой функцией. Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонений.

1.3 Расстояние между объектами (метрика).

Теперь необходимо ввести понятие "расстояние между объектами". Это понятие должно отражать меру сходства, близости объектов между собой по всей совокупности используемых признаков. Иными словами служить интегральной мерой сходства объектов между собой. Расстоянием между объектами в пространстве признаков называется такая величина $d(X_i, X_j)$, которая удовлетворяет следующим аксиомам:

1. $d(X_i, X_j) \geq 0$, äëÿ âñåõ X_i è X_j èç E_z (неотрицательность расстояния);
2. $d(X_i, X_j) = d(X_j, X_i)$ (ñèììåòðèÿ);
3. $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$, ãäå X_i, X_j è X_k ëþáûå òðè âåêòîðà èç E_z (неравенство треугольника)
4. $d(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$ (неразличимость тождественных объектов);

Значение $d(X_i, X_j)$ для заданных X_i и X_j называется расстоянием между X_i и X_j и эквивалентно расстоянию между I_i и I_j соответственно выбранным атрибутам $C= (C_1, C_2, ..., C_z)^T$.

Меру близости (сходства) объектов удобно представить как обратную величину от расстояния между объектами. В многочисленных изданиях посвященных кластерному анализу описано более 50 различных способов вычисления расстояния между объектами. Кроме термина "расстояние" в литературе часто встречается и другой термин - "метрика", который подразумевает метод вычисления того или иного конкретного расстояния. Наиболее распространенный способ в случае количественных признаков является так называемое "евклидово расстояние" или "евклидова метрика".

$$d(X_i, X_j) = \left(\sum_{k=1}^{z} (x_{ki} - x_{kj})^2 \right)^{\frac{1}{2}} \quad (1)$$

Весьма напоминает выражение для евклидового расстояния так называемое обобщенное степенное расстояние Минковского [32], в котором в степенях вместо двойки используется другая величина. В общем случае эта величина обозначается символом "p". При p=2 мы получаем обычное Евклидово расстояние. Ниже приводится выражение для обобщенной метрики Минковского.

$$d(X_i, X_j) = \left(\sum_{k=1}^{z} \left| x_{ki} - x_{kj} \right|^p \right)^{\frac{1}{p}} \quad (2)$$

Используются также такие метрики, как манхэттенское расстояние (расстояние городских кварталов или city-block), метрика "доминирования" (Sup-метрику или Супремум-норма). Метрика Минковского представляет собой большое семейство метрик, включающее и наиболее популярные метрики. Однако существуют методы вычисления расстояния между объектами, принципиально отличающиеся от метрик Минковского. Наиболее важное из них так называемое расстояние Махаланобиса [11, стр. 301-310] .

В работе [20] рассмотрены следующие метрики для работы с номинальными и порядковыми атрибутами: порядковые - метрики Спирмена, Кендалла; номинальные - метрики Жаккара, Рассела-Рао, Бравайса, Юла. Как отмечено ранее, это далеко не все существующие метрики. Соответственно, если у исследователя помимо количественных атрибутов есть атрибуты других типов, то возникает задача найти метрику между разнотипными признаками, измеренными в разных шкалах. В этом случае необходимо решить задачу приведения всех разнотипных шкал к одной общей шкале [26].

1.4 Методы кластеризации.

Выделяют две группы методов *кластерного анализа*: иерархические и неиерархические [17, стр. 257-263]. Иерархические методы предполагают последовательное объединение объектов в кластеры по степени их близости друг к другу или, напротив, последовательное разбиение совокупности объектов на все более мелкие кластеры. В этом случае кластерное решение представляет собой иерархическую структуру вложенных друг в друга кластеров. Неиерархические методы позволяют находить и идентифицировать "сгущения" объектов в пространстве переменных. Выделяют также кластеризацию "с обучением", при которой предполагается, что количество классов известно заранее, и имеется обучающая выборка - набор объектов, для которых известно, к каким классам они принадлежат. Остальные объекты классифицируются по степени их близости к объектам из выборки обучающей.

Основными методами иерархического кластерного анализа являются метод ближнего соседа, метод полной связи, метод средней связи Кинга и метод Уорда (в некоторых изданиях метод Варда). Наиболее универсальным является последний. Существуют также центроидные методы и методы, использующие медиану. В работе [14], описаны отрицательные стороны использования этого типа методов. Неиерархических методов больше, хотя работают они на одних и тех же принципах. По сути, они представляют собой итеративные методы дробления

исходной совокупности. В процессе деления формируются новые кластеры, и так до тех пор, пока не будет выполнено правило остановки. Между собой методы различаются выбором начальной точки, правило формирования новых кластеров и правилом остановки. Чаще всего используется быстрый кластерный анализ, или, как его еще называют, алгоритм *К-средних* [12; 16, стр. 125-134]. При использовании этого алгоритма аналитик заранее фиксирует количество кластеров в результирующем разбиении.

Выбирая между иерархическими и неиерархическими методами, следует обратить внимание на следующие моменты. Неиерархические методы обнаруживают более высокую устойчивость по отношению к выбросам, неверному выбору метрики, включению незначимых переменных в базу для кластеризации и пр. При этом исследователь должен заранее фиксировать результирующее количество кластеров, правило остановки и, если на то есть основания, начальный центр кластера. Последнее условие существенно отражается на эффективности работы алгоритма. Если нет оснований искусственно задать это условие, рекомендуется использовать иерархические методы. Отметим существенное свойство для обеих групп алгоритмов: не всегда правильным решением является кластеризация всех наблюдений. Возможно, наиболее оптимально будет сначала очистить выборку от выбросов, а затем продолжить анализ. Можно также не задавать очень высокий критерий остановки (можно делать остановку, к примеру, когда кластеризовано более 90% наблюдений).

Отметим, что выбор конкретного метода кластеризации, требует от аналитика хорошего знакомства с природой и предпосылками методов, в противном случае полученные результаты будут похожи на «среднюю температуру по больнице». Для того чтобы убедиться в том, что выбранный метод действительно эффективен в данной области, как правило, применяют следующую процедуру: рассматривают несколько априори различных между собой групп и перемешивают их представителей между собой случайным образом. Затем проводят процедуру кластеризации с целью восстановить исходное разбиение на группы. Показателем эффективности работы метода будет доля совпадений объектов в выявленных и исходных группах.

Некоторые авторы приходят к выводу, что выбор метрики и метода кластеризации не является ключевым моментом в кластерном анализе. Такое утверждение возможно в следующих случаях. Во-первых, оно касается только более грубых - неиерархических методов. Во-вторых, в любом случае необходимо выбирать метрику таким образом, чтобы она не противоречила идее выбранного метода объединения кластеров. Особое внимание следует уделить выбору метрики в случае, если переменные являются зависимыми. Адекватная метрика может быть решением данной проблемы.

2 Кластерный анализ биллингов.

2.1 Детализации телефонных переговоров (биллинг).

Дадим определение того, что является биллингом [35, стр. 292]. Биллинг - это зафиксированные параметры событий, произошедших с объектом исследования (ОИ) за определенный период времени. Такими объектами могут быть:

- Телефон, где событием является входящее или исходящее соединение.
- Сотовые телефоны и пейджеры, где событиями могут быть SMS сообщения.
- Банковские счета, где событием является входящий или исходящий перевод денег.
- Интернет соединения, где событием является просто соединение с поставщиком услуги.
- Другие объекты.

Для телефонного номера зафиксированными параметрами события являются, например, дата и время соединения; длительность соединения; номер телефона, с которым произошло соединение и т.п. Для банковских счетов - дата и время перевода, сумма перевода, направление перевода и т.п.

В качестве примера биллинга рассмотрим детализацию телефонных переговоров. Детализации телефонных переговоров имеют в зависимости от компании-автора, в которой ведется учет соединений, разные структуры. Эти структуры постоянно обновляются и добавляются, в них часто вводят сложные правила и зависимости, из-за чего практически невозможно создать инструментарий, позволяющий автоматически обрабатывать все возможные типы структур. Приведем пример детализации телефонных переговоров:

Звонки с номера 0959222301 за интервал с 01.01.2002 00:00:01 по 06.08.2002 23:59:59

тел. - 0959222301 imsi - Амирасланов Азер Агали оглы лиц. счет - 1746762 Частное лицо

Mob_num	Num	Dur	Dir	Num_A	Cell_Id	Bts_addr
0959222301	+70951079565	01.01.02	13:30:26	168	O	Неизвестно
0959222301	+70951713495	01.01.02	14:55:12	3	O	Неизвестно
0959222301	+70951713495	01.01.02	14:55:50	40	O	Неизвестно

Опишем формально детализацию телефонных переговоров. Заголовочная часть *Z* детализации в общем случае можно представить следующим образом:

$$Z\ (N_1, D_1, D_2, F_1), \text{ где} \qquad (3)$$

N_1 – телефонный номер детализации
D_1 – дата начала детализации
D_2 – дата окончания периода детализации
F_1 – ФИО лица, которому принадлежит телефонный номер детализации

Строка *S* детализации в общем случае может быть представлена следующими атрибутами:

$$S\ (\text{Num}_1, D_1, \text{Dlit}, \text{Nap}), \text{ где} \qquad (4)$$

$\text{Num}_1 = \{n_1, n_2, \ldots, n_r\}$ - все неповторяющиеся номера телефонов детализации, на или с которых произошло соединение с N_1,
r – количество неповторяющихся номеров телефонов в детализации,
D_1 – дата соединения,
$\text{Dlit} = \{\text{dlit}_1, \text{dlit}_2, \ldots, \text{dlit}_f\}$ – длительность соединения,
f – количество строк в детализации,
$\text{Nap} = \{\text{nap}_1, \text{nap}_2, \ldots, \text{nap}_f\} = \{\text{исх}, \text{вх}\}$ – направление соединения (входящее или исходящее соединение).

В работе [35, стр. 292] описан механизм загрузки биллингов в базе знаний (БЗ) системы «Аналитик» в виде расширенных семантических сетей (РСС) [28; 34]. После загрузки биллингов у аналитика появляется возможность воспользоваться аналитическими режимами обработки накопленной информации. Для аналитической обработки биллингов в системе «Аналитик» реализован режим «Детализация номерных объектов» [33], который можно условно разделить на четыре подрежима:

- Граф телефонных переговоров
- Диаграмма длительности переговоров
- Граф финансовых потоков
- Диаграмма финансовых потоков

В зависимости от режима анализа информация выдается на экран двумя способами отображения: графом и диаграммой длительности телефонных переговоров и банковских переводов.

В первом режиме задача отображения решена встроенными средствами языка Декл [40]. На основе анализа, проведенного по всей БЗ, выделяются все детализации по выбранному телефонному номеру. Будем называть такой телефонный номер объектом исследования.

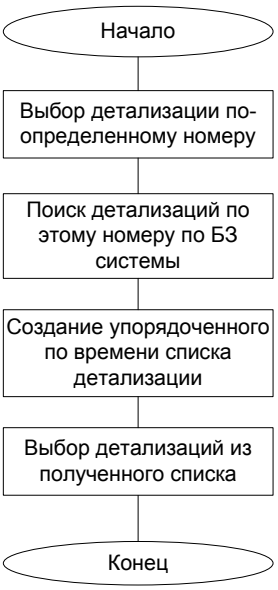

Рис. 1. Блок-схема алгоритма поиска подобных детализаций.

Таким образом, пользователь имеет возможность анализировать активность ОИ в любые промежутки времени, за которые в БЗ загружены детализации по этому телефону, используя алгоритм, приведенный на рис. 1. Затем проводится определение особо активных телефонов, с которыми ОИ разговаривал либо чаще всего, либо дольше всего, т.е. ведется подсчет количества входящих и исходящих звонков между ОИ и другими телефонами за выбранный промежуток времени и подсчет длительности переговоров. Затем в зависимости от установленных пользователем интервалов длительности и частоты соединений телефонам присваивается определенный вес, т.е. для каждой строки детализации $m=\{1, ..., f\}$ и каждого телефонного номера n_p из Num_1, где $p=\{1, ..., r\}$: при условии, что $\mathrm{nap_m}$ = исх вычисляем $S_{1p} = \sum_{i=1}^{k_p} dlit_m$, где k_p – количество строк детализации, удовлетворяющих этим условиям; при условии, что $\mathrm{nap_m}$ = *вх* вычисляем $S_{2p} = \sum_{i=1}^{l_p} dlit_m$, где l_p – количество строк детализации, удовлетворяющих этому условию.

В результате получается следующий набор атрибутов результирующей агрегированной информации *Itog*:

$Itog\ (Num_1, S_1, S_2, K_1, K_2)$, где (5)

$\mathrm{Num}_1 = \{n_1, n_2, ..., n_r\}$ – множество неповторяющихся номеров телефонов,
$S_1 = \{ S_{11}, S_{12}, ..., S_{1p} \}$ – сумма длительностей исходящих соединений,
$S_2 = \{ S_{21}, S_{22}, ..., S_{2p} \}$ – сумма длительностей входящих соединений,
$K_1 = \{k_1, k_2, ..., k_p\}$ – количество исходящих соединений,
$K_2 = \{l_1, l_2, ..., l_p \}$– количество входящих соединений, $p = \{1..r\}$

На основании полученной информации создается графический объект со следующими атрибутами: номер телефона, суммарная длительность исходящих и/или входящих соединений с рассматриваемым телефоном, вес длительности, количество входящих и/или исходящих звонков, вес суммарного числа звонков. Затем все созданные объекты визуализируются на графе (Рис. 2). В зависимости от весов информация для наглядности раскрашивается пятью разными цветами. Телефоны можно отображать как в виде окружности, так и в виде дерева.

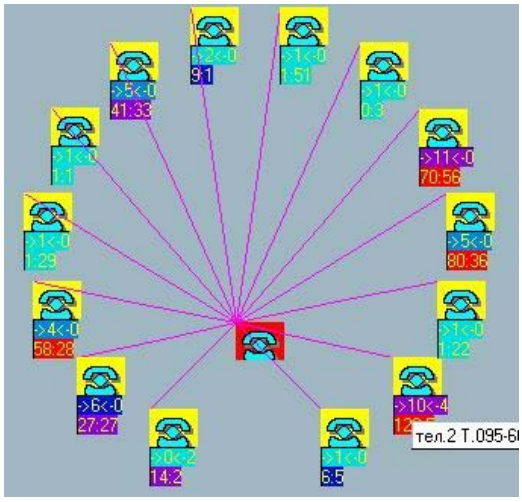

Рис. 2. Визуализация детализации телефонных переговоров на графе.

Для расширения возможностей аналитиков, работающих с системой «Аналитик», рассмотрим возможность применения кластерного анализа к детализациям телефонных переговоров.

2.2 Объект кластеризации.

Рассмотрим в качестве примера применения кластерного анализа кластеризацию детализация телефонных переговоров. В качестве объектов кластеризации будем рассматривать телефонные номера, которые попали в детализацию по определенному телефону (см. пример 1).

В предыдущем разделе была описана структура детализации, состоящая только из количественных атрибутов, что существенно облегчает кластеризацию. Отметим также взаимозависимости атрибутов детализаций – поскольку суммарная длительность соединений не может быть больше, чем длительность рассматриваемого периода (дня, месяца и т.п.), то чем больше соединений произошло за этот период, тем меньше усредненная длительность соединения. Эта зависимость является косвенной, поэтому не должна сказаться на результатах кластеризации. Таким образом, нет необходимости предварительной классификации нечисловых атрибутов и поиска взаимозависимых параметров при выборе адекватной метрики для определения расстояний между объектами. В кластеризации не участвует атрибут Num_1, т.к. он является идентификатором объектов анализа.

Для определения оптимальных критериев кластеризации биллингов проведем следующий эксперимент. В качестве исходных данных[1] рассмотрим такой экземпляр Itog (Табл. 1), который изначально можно разбить на определенное число кластеров. Перемешаем «наблюдения» между собой так, чтобы максимально усложнить задачу кластеризации. Применим к этим данным различные комбинации метрик и алгоритмов кластеризации, как это описано в пункте 1.4. Та комбинация, результатом работы которой будет разбиение, максимально похожее на изначальное, будет оптимальной для кластеризации биллингов.

Таблица 1. Исходные данные кластеризации.

Номер телефона (Num_1)	**Длительность исходящих соединений** (S_1)	**Длительность входящих соединений** (S_2)	**Количество исходящих соединений** (K_1)	**Количество входящих соединений** (K_2)
111	121	1200	5	112
222	878	200	99	4
333	21	600	1	10
444	500	300	60	12
555	154	978	8	88
666	2500	232	2314	29
777	1400	400	130	20

[1] Поскольку реальные детализации на существующие телефоны являются закрытой информацией, которую могут получать либо владельцы этих телефонов, либо силовые ведомства, воспользуемся синтетической информацией.

888	113	1144	1	70
999	2134	122	1700	18
900	11	578	7	16

Из таблицы 2 несложно выделить 4 группы – кластера – телефонов Num_1:

1) π_1 = {111 (121, 1200, 5, 112), 555 (154, 978, 8, 88), 888 (113, 1144, 1, 70)}
2) π_2 = {222 (878, 200, 99, 4), 444 (500, 300, 60, 12), 777 (1400, 400, 130, 20)}
3) π_3 = {666 (2500, 232, 2314, 29), 999 (2134, 122, 1700, 18)}
4) π_4 = {333 (21, 600, 1, 10), 900 (11, 578, 7, 16)}

Кроме схожих количественных показателей, критерием такой группировки являются интегрированные показатели средней длительности входящих и исходящих соединений в группе. Среднюю длительность входящих и исходящих для каждого элемента в группе вычисляется по формулам $d_{âõ} = S_2/K_2$ и $d_{èñõ} = S_1/K_1$. Усредненная длительность соединений представлена в табл. 2.

Таблица 2. Усредненная длительность соединений.

Номер телефона (Num_1)	**Средняя длительность исходящих** ($d_{èñõ}$)	**Средняя длительность входящих** ($d_{âõ}$)	**Количество исходящих соединений** (K_1)	**Количество входящих соединений** (K_2)
111	24,2	10,7	5	112
222	8,89	50	99	4
333	21	60	1	10
444	8,3	25	60	12
555	19,25	11,1	8	88
666	1,08	8	2314	29
777	10,77	20	130	20
888	113	16,3	1	70
999	1,26	6,8	1700	18
900	1,5	36	7	16

Средняя длительность исходящих и входящих соединений в группе рассчитываются соответственно по формулам:

$\overline{d}_{èñõ} = \frac{\sum_{i=1}^{n} S_{1i}}{\sum_{i=1}^{n} k_i}$, $\overline{d}_{âõ} = \frac{\sum_{i=1}^{n} S_{2i}}{\sum_{i=1}^{n} l_i}$, где n – количество элементов в группе, i – номер элемента в группе.

Среднее количество исходящих и входящих соединений рассчитывается соответственно по формулам:

$\overline{k}_{èñõ} = \frac{\sum_{i=1}^{n} k_i}{n}$; $\overline{k}_{âõ} = \frac{\sum_{i=1}^{n} l_i}{n}$, где n – количество элементов в группе, i – номер элемента в группе. Усредненные показатели кластеров представлены в табл. 3.

Таблица 3. Усредненные показатели групп.

Номер группы	**Средняя длительность исходящих по группе** ($\overline{d}_{èñõ}$)	**Средняя длительность входящих по группе** ($\overline{d}_{âõ}$)	**Среднее количество исходящих** ($\overline{k}_{èñõ}$)	**Среднее количество входящих** ($\overline{k}_{âõ}$)
1	52,15	12,7	4,7	90
2	27,96	31,7	96,3	12
3	1,17	7,4	2007	23,5

4	11,25	48	4	13

Таким образом, по каждой группе - кластеру - можно сделать следующие выводы:

1) Кластер π_1 характеризует большое количество *входящих* со средней длительностью 12 минут, небольшое количество исходящих высокой длительности 52 минуты.
2) Кластер π_2 характеризует большое количество *исходящих* со средней длительностью 28 минут, небольшое количество входящих средней длительности 32 минуты.
3) Кластер π_3 характеризует очень большое количество *исходящих* соединений (2007 шт.) малой длительности 1 минута, небольшое число входящих малой длительности 7 минут.
4) Кластер π_4 характеризует небольшое число *входящих* с большой длительностью 48 минут, небольшое количество исходящих средней длительности 11 минут.

Для сравнения результатов кластеризации различными алгоритмами кластерного анализа с оптимальным разбиением, воспользуемся целевой функцией – внутригрупповой суммой квадратов отклонений (СКО). Рассчитаем для каждой группы СКО по следующей формуле:

$$W = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2$$

Для расчета воспользуемся программным пакетом Maple 8 [23]. СКО рассчитывается для каждого признака каждого кластера. Соответственно, для первого кластера π_1 будет рассчитано 4 СКО (в соответствии с количеством атрибутов объектов кластеризации):

$W_1=W_{11}+W_{12}+W_{13}+W_{14}$ – суммарное СКО первого кластера

В расчете каждого СКО будет принимать участие три (в соответствии с количеством объектов в кластере π_1) значения первого атрибута каждого из объектов кластера π_1 (S_{11}=121, S_{15}=154, S_{18}=113). Приведем пример расчета первого СКО:

$$W_{11}=121^2+154^2+113^2-\frac{1}{3}(121+154+113)^2=51126-\frac{1}{3}(388)^2=51126-50181.3=944.7$$

Так же рассчитывается СКО для второго, третьего и четвертого признаков объектов первого кластера. Получаем:

W_1=944.7+26658.7+24.7+888=28516.1

Рассчитаем также W_2, W_3, W_4.

W_2=408456+20000+2460.6666666666+128=431044.7
W_3=66978+6050+188498+60.5=261586.5
W_4=50+242+18+18=328

Таким образом, сумма квадратов отклонений для данной кластеризации W равно

$W=W_1+W_2+W_3+W_4$=28516.1+431044.7+261586.5+328=721475.3

2.3 Выбор метрики и расчет расстояний между объектами.

Поскольку исходные данные кластеризации являются численными, воспользуемся обычной в таких случаях евклидовой метрикой[1] (см. п. 1.4):

[1] Евклидовой метрикой воспользуемся для расчета «вручную» расстояний между объектами. В дальнейшем расчет расстояние между объектами, как и сама кластеризация, будет проводиться автоматически, при помощи соответствующего ПО.

$$d(X_i, X_j) = (\sum_{k=1}^{z} (x_{ki} - x_{kj})^2)^{\frac{1}{2}}$$

В нашем случае *z*, количество атрибутов объектов, равно четырем. Рассчитаем расстояние между объектами (Табл. 4), представленными в таблице 2. Для этого того вычислим в начале $d^2(X_i, X_j)$ по формуле:

$$d^2(X_i, X_j) = \sum_{k=1}^{4} (x_{ki} - x_{kj})^2$$

Таблица 4. Квадрат расстояний между объектами.

№	1	2	3	4	5	6	7	8	9	10
1	0	-	-	-	-	-	-	-	-	-
2	1593549	0	-	-	-	-	-	-	-	-
3	380420	904089	0	-	-	-	-	-	-	-
4	966666	154469	322926	0	-	-	-	-	-	-
5	50958	1144797	166706	587880	0	-	-	-	-	-
6	11935035	7538758	11631195	9085429	11381349	0	-	-	-	-
7	2299930	313701	1958382	824964	1906108	6008161	0	-	-	-
8	4980	1490321	308000	868950	29610	11881163	2229046	0	-	-
9	8096114	4147017	7579918	5391276	7520900	523173	3080944	8018230	0	-
10	408204	903181	656	319230	185634	11637255	1976150	333712	7581318	0

Учитывая одно из свойств расстояния – симметричность (см. п. 1.3), в соответствии с которым расстояния между объектами 1-2 и 2-1 равно, рассчитывать расстояние только для одной из пар. Расстояния симметричны относительно нулевой диагонали. Теперь рассчитаем $d(X_i, X_j)$ (Табл. 5).

Таблица 5. Матрица расстояний между объектами.

№	1	2	3	4	5	6	7	8	9	10
1	0	-	-	-	-	-	-	-	-	-
2	1262.359	0	-	-	-	-	-	-	-	-
3	616.782	950.836	0	-	-	-	-	-	-	-
4	983.1917	393.025	568.2658	0	-	-	-	-	-	-
5	225.7388	1069.95	408.2965	766.733	0	-	-	-	-	-
6	3454.712	2745.68	3410.454	3014.2	3373.625	0	-	-	-	-
7	1516.552	560.09	1399.422	908.275	1380.619	2451.155	0	-	-	-
8	70.56912	1220.79	554.9775	932.175	172.0756	3446.906	1493	0	-	-
9	2845.367	2036.42	2753.165	2321.91	2742.426	723.307	1755.26	2831.648	0	-
10	638.9084	950.358	25.6125	565.004	430.8526	3411.342	1405.76	577.6781	2753.42	0

На основании данных из табл. 5 проведем с помощью программы Attestat [21] кластеризацию тремя различными способами: методами средней связи Кинга, Уорда, k-средних Мак-Куина. Для каждого из способов в качестве исходных данных кластеризации укажем предполагаемое количество кластеров *m*=4, на которое необходимо разбить множество объектов кластеризации.

Полученные методом средней связи Кинга (иерархическая кластеризация) результаты представлены в табл. 6.

Таблица 6. Результаты кластеризации методом средней связи Кинга.

№ кластера	Количество объектов в кластере	№ объекта	№ объекта	№ объекта	№ объекта	№ объекта

π_1	5	1	8	5	3	10
π_2	2	2	4			
π_3	2	6	9			
π_4	1	7				

Последовательность объединений кластеров представлена в табл. 7.

Таблица 7. Последовательность объединения объектов.

Объединенные объекты		**Уровень связи**
3	10	25.6125
1	8	70.56912
1	5	197.5829
2	4	393.0254
1	3	506.2912
3	5	723.307
2	4	728.5021
1	2	1098.596
1	2	2656.911

Отметим, что при объединении двух объектов в одни кластер, расстояние между кластерами в процессе кластеризации пересчитывается. Рассчитаем СКО.

W_1=16408+349304+43.2+8020.8=373776
W_2=71442+5000+760.5+32=77234.5
W_3=66978+6050+188498+60.5=261586.5
W_4=0

Таким образом, сумма квадратов отклонений W для данного разбиения равно:

$W=W_1+W_2+W_3+W_4$=373776+77234.5+261586.5+0=712597

Результаты работы метода Уорда (иерархическая кластеризация) представлены в табл. 8. Дендограмма последовательности объединения объектов представлена на рис. 3.

Таблица 8. Результаты кластеризации методом Уорда.

№ кластера	**Количество объектов в кластере**	**№ объекта**	**№ объекта**	**№ объекта**
π_1	3	1	8	5
π_2	3	2	4	7
π_3	2	3	10	
π_4	2	6	9	

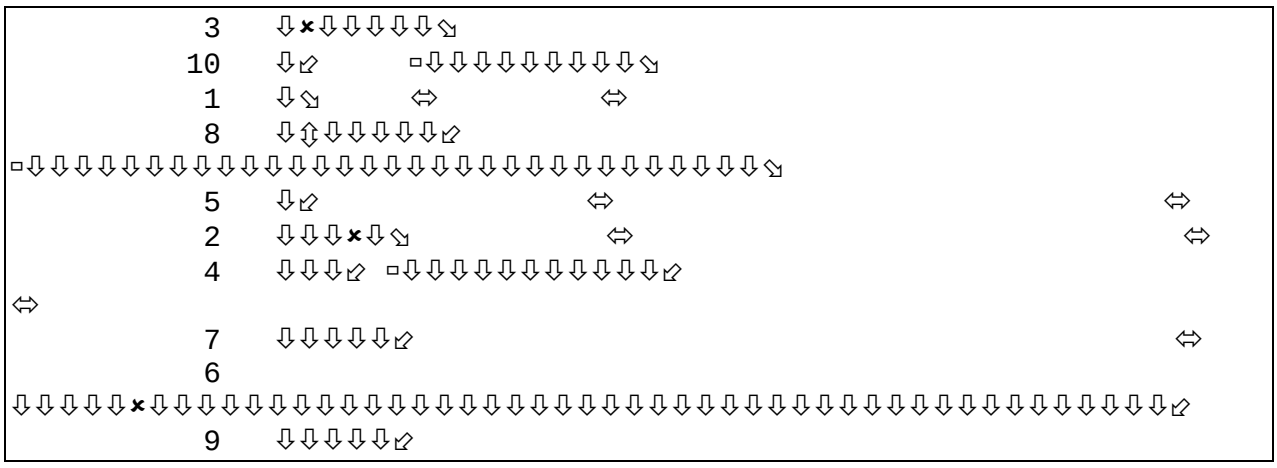

Рис. 3. Дендограмма последовательности объединения объектов

В этом случае разбиение и, соответственно, СКО совпадает с оптимальным разбиением и его СКО.

Результатом работа метода k-средних Мак-Куина (неиерархическая кластеризация) является разбиение, представленное в табл.9. Кординаты центров тяжести кластеров представлены в табл. 10.

Таблица 9. Результаты кластеризации методом k-средних Мак-Куина.

№ кластера	**Количество объектов в кластере**	**№ объекта**	**№ объекта**	**№ объекта**	**№ объекта**	**№ объекта**
π_1	5	1	3	5	8	10
π_2	0					
π_3	2	2	4			
π_4	3	6	7	9		

Таблица 10 Центры тяжести кластеров.

№ кластера	**Длительность исходящих соединений (Sum_1)**	**Длительность входящих соединений (Sum_2)**	**Количество исходящих соединений (K_1)**	**Количество входящих соединений (K_2)**
π_1	84	900	4.4	59.2
π_2	0	0	0	0
π_3	689	250	79.5	8
π_4	2011.333	251.3333	1381.333	22.33333

В результатах разбиения методом k-средних Мак-Куина можно отметить, что кластер π_2 не содержит объектов. Если в качестве исходных данных кластеризации методом k-средних указать целевое количество кластеров m=3, то кластеры π_1, π_2, π_3 содержат те же объекты, что и π_1, π_2, π_4 при текущем разбиении.

Рассчитаем СКО.

W_1=16408+349304+43.2+8020.8=373776
W_2=0
W_3=71442+5000+760.5+32=77234.5
W_4=627570.7+39202.7+2537250.7+68.7=3204092.8

Таким образом, получим сумму W квадратов отклонений для данной кластеризации

$W=W_1+W_2+W_3+W_4$=373776+0+77234.5+3204092.8=3655103.3

Проведем кластеризацию заданных объектов с применением различных комбинаций метрик в программных комплексах SPSS и Statgraphics. Результаты кластеризации представлены в таблице 11. В первых десяти строках представлены объекты и результаты их кластеризации. Курсивом выделены результаты, совпадающие с результатами исходного разбиения. В нижней части таблицы представлены методы кластеризации (использовались как иерархические методы, так и неиерархические), метрики и программы кластеризации. В качестве исходных данных задавалось количество кластеров m=4, на которое разбивалось множество объектов кластеризации.

Таблица 11. Сводная таблица кластеризации

№ группы \ № объекта	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	*1*	*1*	*1*	*1*	1	1	1	1	1	**1**	**1**
2	*2*	*2*	*2*	*2*	4	4	2	2	2	**2**	**2**
3	*3*	*3*	*3*	*3*	1	1	3	1	1	**1**	**1**
4	*2*	*2*	*2*	*2*	4	4	2	2	2	**2**	**2**
5	*1*	*1*	*1*	*1*	1	1	1	1	1	**1**	**1**
6	*4*	*4*	*4*	*4*	2	3	4	3	3	**3**	**3**
7	*2*	*2*	*2*	*2*	4	3	3	2	2	**4**	**4**
8	*1*	*1*	*1*	*1*	1	1	1	1	1	**1**	**1**
9	*4*	*4*	*4*	*4*	3	3	4	4	4	**3**	**3**
10	*3*	*3*	*3*	*3*	1	1	3	1	1	**1**	**1**
Метод	Уорд	Уорд	Уорд	k-средних	k-средних	k-средних	k-средних	медианы	средней связи	Уорд, медианы	средней связи
Метрика	сити-блок, минковского	Махаланобис, эвклидова, сити-блок	сити-блок, эвклидова	сити-блок,	эвклидова	эвклидова	эвклидова	ближнего соседа, дальнего соседа, центроидный, сити-блок	сити-блок	эвклидова	эвклидова
Программа	SPSS	Attestat	STAT.	STAT.	SPSS	Attestat	STAT.	SPSS	Attestat, SPSS	SPSS	Attestat, SPSS
СКО	*721476*	*721476*	*721476*	*721476*	804821	3655103	1679066	804820.6	804821	**712597**	**712597**

3 Заключение.

На основании полученных результатов можно сделать следующие выводы:

1) Наименьшее СКО дали иерархические методы кластеризации (метод Уорда, средней связи и медианы).
2) Метод Уорда дал тот же результат, что был получен до кластеризации.
3) СКО кластеризации в группах исследований №10 и №11 (Табл. 11) составляет 712597, что меньше, чем СКО исходного разбиения. В обеих группах использовались иерархические методы кластеризации и Эвклидова метрика. Недостатком этой группы является неравномерность разбиения. Так в кластер π_1 помещено 5 объектов, в кластер π_2 – 2, в кластер π_3 – 3, в кластер π_4 – 1. У кластера π_4 СКО равно нулю, кластер π_1 имеет СКО равное 373 776, что значительно больше СКО кластера π_1 при разбиении, сделанным до кластеризации (28516). Таким образом, данное разбиение не является оптимальным с аналитической точки зрения, при этом являясь оптимальным с точки зрения СКО.
4) Различные комбинации метрик и способов кластеризации дают различные результаты, т.е. кластеры содержат различные объекты, соответственно, их разбиение дает разное СКО.
5) Одна и та же пара алгоритма кластеризации и метрики дает различные результаты в зависимости от программы кластеризации. Так метод k-means и Эвклидова метрика, в зависимости от программного пакета, дают три разных разбиения и, соответственно, три разных СКО. Это говорит о том, что один и тот же алгоритм или метрика могут быть по-разному проинтерпретированы разработчиками программных статистических комплексов.
6) Время кластеризации суммируется из времени, необходимого для группировки всех звонков в биллинге по телефонному номеру, и времени кластеризации сгруппированных данных. Исследования показали, что время, необходимое на группировку, больше, чем время кластеризации результатов группировки. Если для группировки использовать современные СУБД, это соотношение может быть изменено. В любом случае, время кластеризации не является значительным и исчисляется несколькими секундами.

Суммируя вышесказанное можно сказать, что кластерный анализ может быть эффективно использован в аналитических задачах в различных предметных областях, где используются биллинги. Он позволит существенно снизить трудозатраты аналитиков в процессе обработки значительных массивов информации и решить задачи классификации и выделения похожих групп объектов. Результаты исследования могут быть использованы в решении таких важных задач, как: выявлении преступных групп лиц в криминалистике, разработке новых тарифов и предложений в мобильной связи, выявлении групп счетов в банковском секторе и т.п.

Одной из причин, по которой данный вид анализа в применении к детализациям телефонных переговоров еще не исследован, заключается в том, что детализации во всех биллинговых системах являются закрытой информацией. Эту информацию могут получить либо абоненты биллинговых систем по своим собственным услугам, либо силовые ведомства в рамках следственных мероприятий. Таким образом «Аналитик» на сегодняшний день является одной из немногих логико-аналитических систем, имеющих в качестве исходных данных детализации и несколько режимов для их обработки. Реализация кластерного анализа детализаций существенно расширит аналитические возможности системы.

Приложение 1

Сравнительный анализ методов кластеризации.

На основе алгоритма кластеризации описанного в работе [7, стр. 134-138], выделим следующие основные этапы кластерного анализа:

1. Изменение исходных данных:
   - Выбор метрики
   - Выбор метода стандартизации
   - Определение взаимосвязанных объектов в выборке
2. Принятие решений:
   - Определение количества кластеров, на которое необходимо провести разбиение
   - Выбор метода кластеризации

- Исключение подвыборок, если это необходимо

3. Анализ полученных результатов, состоящий из ответов на следующие вопросы:
   - Насколько полученное разбиение отличается от случайного?
   - Является ли оно надежным и стабильным на подвыборках?
   - Какова взаимосвязь между результатами кластеризации и переменными, не участвовавшими в процессе кластеризации?
   - Можно ли проинтерпретировать полученные результаты?
   - По какому набору переменных проводить кластеризацию наиболее эффективно?

Необходимо обратить внимание на то, что, в общем случае, все эти этапы взаимосвязаны, и решения, принятые на каждом из них взаимообуславливают друг друга.

Литература.

1. Ball, G.H., 'Data-analysis in the social sciences: What about the details?', Proceedings of the Fall Joint Computer Conference, 27,533-559 (1966)
2. Cole A.J., Numerical Taxonomy, Academic Press, New York (1969).;
3. Cormack, R.M., 'A review of classification', Journal of the Royal Statistical Society, Series A, 134, 321-353 (1971)
4. Crouch, D., 'A clustering algorithm for large and dynamic document collections', Ph.D. Thesis, Southern Methodist University. Dallas. 1972.
5. Dorofeyuk, A.A., 'Automatic Classification Algorithms (Review)', Automation and Remote Control, 32,1928-1958 (1971).
6. Fritzche, M., 'Automatic clustering techniques in information retrieval' Diplomarbeit, Institut fur Informatik der Universitat Stuttgart (1973)
7. Punj G., Stewart D. Кластерный анализ в маркетинговых исследованиях: обзор и предпосылки применения. Journal of Marketing Research, Vol. XX, (May 1983), pp.134-148
8. Good, I.J., 'Categorization of classification' In Mathematics and Computer Science in Biology and Medicine, HMSO, London, 115-125 (1965).
9. Hartigan, J.A. (1975), Clustering Algorithms, NY: Wiley
10. Litofsky B., Utility of automatic classification systems for information storage and retrieval, Ph.D. Thesis, University of Pennsylvania. Philadelphia. 1969.
11. Mahalanobis P.C. Analysis of race mixture in Bengal, J. Asiat. Soc. (India), Vol. 23, (1925), 301-310
12. Murtagh, F., Multidimensional clustering algorithms, Compstat Lectures, Heidelberg: Physica-Verlag, 1985
13. Prywes, N.S. and Smith, D.P., 'Organization of Information', Annual Review of Information Science and Technology, 7,103-158 (1972).
14. Sneath, P.H.A. and Sokal, R.R., Numerical Taxonomy: The Principles and Practice of Numerical Classification, W.H. Freeman and Company, San Francisco (1973)
15. Tryon R.C. Cluster Analysis. New York: McGraw-Hill. - 1939
16. Wishart, D. (1986), Estimation of Missing Values and Diagnosis Using Hierarchical Classifications, Computational Statistics Quarterly, 2(1), p. 125-134
17. Wishart D. “Exploiting the graphical user interface in statistical software: the next generation”. *Interface '98*. Computing Science and Statistics, 30, 1998, 257-263
18. С. А. Айвазян, В. С. Степанов. Инструменты статистического анализа данных. Мир ПК, №08 – М.: Издательство «Открытые системы». Москва. 1997 г.
19. Боровиков В. STATISTICA: искусство анализа данных на компьютере (с CD-ROM), 2 издание. Питер. 2003
20. Гайдышев И.П. Анализ и обработка данных. Специальный справочник. – СПб.: Издательство «Питер», 2001 г., 752 стр.
21. Гайдышев И.П. Решение научных и инженерных задач средствами Excel, VBA и C++ (+ CD). – СПб.: Издательство «БХВ-Петербург», 2004 г., 512 стр
22. Григорьев С.Г., Левандовский В.В., Перфилов А.М., Юнкеров А.И. Пакет прикладных программ Statgraphics на персональном компьютере. Практическое пособие по обработке результатов медико-биологических исследований. СПб., 1992
23. Maple 8 в математике, физике и образовании. / В.П. Дьяконов – М.: СОЛОН-Пресс, 2003. 656 стр.
24. Дюран Б., Оделл П. Кластерный анализ. - М.: Статистика, - 1977, - 128 с.

25. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа: пакет ППСА. - М.: Финансы и статистика, 1986. - 232с
26. Загоруйко Н.Г., Елкина В.Н., Лбов Г.С. Алгоритмы обнаружения эмпирических закономерностей. – Новосибирск: Наука, 1985
27. Статистический словарь/ гл.ред. М.А. Королёв. - М.: Финансы и статистика, 1989 г.- 623 с.
28. Кузнецов И.П., Семантические представления. Отв. ред. Е. В. Золотов – М:. Наука 1986
29. Kuznetsov Igor, Matskevich Andrey. System for Extracting Semantic Information from Natural Language Text. Труды международного семинара Диалог-2002 по компьютерной лингвистике и ее приложениям. Том 2. Протвино. – М.: Издательство «Наука», 2002.
30. Кулаичев А.П. Методы и средства анализа данных в среде Windows. STADIA 6.0. – М.: Информатика и компьютеры, 1998. 270 с.
31. Луценко Е.В. Теоретические основы и технология адаптивного семантического анализа в поддержке принятия решений (на примере универсальной автоматизированной системы распознавания образов "ЭЙДОС-5.1"). - Краснодар: КЮИ МВД РФ, 1996. - 280с.
32. Мандель И.Д. Кластерный анализ. - М.: Финансы и статистика. 1988. - 176с.
33. Рабинович Б.И., Гнидо Е.И., Кузнецов И.П., Мацкевич А.Г. Частотный анализ биллингов телефонных переговоров в Логико-Аналитической системе «Аналитик». Научно-техническая конференция профессорско-преподавательского, научного и инженерно-технического состава. МТУСИ. Москва, 29-31 января 2002 г. Тезисы докладов. Издательство ООО «Инсвязьиздат». Москва 2002 г
34. Рабинович Б.И. Аналитическая система обработки и управления структурированной информацией. Интеллектуальные технологии и система. Выпуск 5. – М.: Издательство ООО «Эликс+», 2003, стр. 284-296.
35. Рабинович Б.И. Система сбора и обработки разнородной информации «АНАЛИТИК». Интеллектуальные технологии и система. Выпуск 7. – М.: Издательство ООО «Эликс+», 2005, стр. 284-296.
36. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. М.: МедиаСфера, 2003. 312 с
37. Симанков В.С., Луценко Е.В. Адаптивное управление сложными системами на основе теории распознавания образов. Монография (научное издание). – Краснодар: ТУ КубГТУ, 1999. - 318с
38. Статистические и математические системы // Каталог «Тысячи программных продуктов». – 1995. - №2. М.
39. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М: ИНФРА-М. Финансы и статистика, 1995. 384 с.
40. Шарнин М.М, Кузнецов И.П. Продукционный язык программирования ДЕКЛ. В сб. «Система обработки декларативных структур знаний Деклар-2». ИПИАН/СССР - М.: Изд-во «Наука», 1988 – стр. 134-152
41. Кластерный анализ: основы метода и его применение в биомедицине [Электронный ресурс]/ Статья. Леонов В.П. - Режим доступа: http://www.biometrica.tomsk.ru/cluster.htm - Загл. с экрана. - Яз. рус., англ.
42. Кластерный анализ [Электронный ресурс]/ StatSoft - Режим доступа: http://www.statsoft.ru/home/textbook/modules/stcluan.html#general - Загл. с экрана. - Яз. рус., англ.