Московский государственный университет путей сообщения
(МИИТ)

Кафедра «Автоматизированные системы управления»

А.В. КУТЫРКИН, А.В. СЁМИН

КЛАСТЕРНЫЙ АНАЛИЗ

Методические указания к лабораторной работе
по дисциплинам «Системы искусственного интеллекта», «Представление и обработка знаний» для студентов специальностей
«Информационные системы и технологии» и
«Автоматизированные системы обработки информации и управления»

МОСКВА 2005

УДК 519.237.8
К95

Кутыркин А.В., Сёмин А.В. Кластерный анализ: Методические указания. — М.: МИИТ, 2005, — 22 с.

Методические указания посвящены вопросам разработки и исследования моделей и методов кластерного анализа при интеллектуальной обработке данных в информационных системах. Представлено описание основных алгоритмов построения минимального остовного дерева Крускала и Прима, которые предлагается использовать при построении эффективного алгоритма кластерного анализа и его программной реализации в среде Borland Delphi. Включены рабочее задание и контрольные вопросы.

СОДЕРЖАНИЕ

ЦЕЛЬ РАБОТЫ: ознакомление с проблемой кластерного анализа при интеллектуальной обработке данных в информационных системах; изучение алгоритмов кластеризации, использующих построение минимального остовного дерева; приобретение навыков в программной реализации изученных алгоритмов в среде Borland Delphi и в компьютерном проведении кластерного анализа.

1. ОБЩИЕ СВЕДЕНИЯ О КЛАСТЕРНОМ АНАЛИЗЕ

Хорошо известно, что новые знания о предметной области (ПО) управления лежат в основе принятия эффективных революционных решений во всех сферах организационного и технического управления. Возможность получения новых знаний, путём извлечения полезной информации из совокупности данных, описывающих ПО управления, представляет собой существенное достижение современных информационных технологий.

В последнее время одним из самых мощных инструментариев, помогающим извлечь из различных, в том числе и больших, баз данных ранее неизвестные знания о ПО управления, являются средства интеллектуального анализа данных (ИАД) Data Mining (дословно – добыча данных). Средства Data Mining, называемые также Knowledge Discovery In Data (обнаружение знаний в данных), позволяют существенно расширить круг практически значимых задач управления, решаемых с использованием компьютеров. Применение ИАД стало в настоящее время частью экономической стратегии многих компаний.

Обнаружение новых знаний с помощью ИАД осуществляется с помощью широкого набора инструментальных средств, среди которых важное место занимает кластерный анализ.

Задача кластерного анализа заключается в выявлении естественного локального сгущения объектов, каждый из которых описан набором переменных или признаков. В процессе кластерного анализа осуществляется разбиение исследуемого множества объектов, представленных

многомерными данными, на группы похожих в определённом смысле объектов, называемых кластерами.

Слово кластер английского происхождения (cluster) и переводится как сгусток, пучок, группа объектов, характеризуемых общими свойствами. Родственными понятиями, используемыми в литературе вместо понятия кластер, являются – класс, таксон, страта, сегмент. Поэтому для задачи кластерного анализа могут также употребляться и следующие термины: автоматическая классификация, обучение без учителя, самообучение, таксономия, стратификация, сегментация.

Кластерный анализ лежит в основе любой интеллектуальной деятельности и является фундаментальным процессом в науке. Любые факты и явления должны быть упорядочены или сгруппированы по их схожести, т.е. классифицированы, прежде чем разрабатываются общие принципы, объясняющие их поведение и взаимную связь. Необходимость классификации признавалась ещё Аристотелем. Ярким примером удачно выполненного кластерного анализа является открытая Д.И. Менделеевым периодическая система элементов.

Кластерный анализ может быть применён к любой предметной области, где необходимо исследовать объекты, заданные экспериментальными или статистическими данными. Применение кластерного анализа не требует предварительных знаний об анализируемых данных, что позволяет его использовать для данных практически произвольной природы. Поэтому задача кластерного анализа обычно решается на начальных этапах исследования, когда о данных мало чего известно. Её решение помогает лучше понять природу анализируемых объектов.

Большая практическая ценность кластерного анализа заключается в том, что он может производить группировку объектов не только по одному параметру, но и по целому набору признаков. Это открывает широкие возможности для проведения кластерного анализа записей в хранилищах и

базах данных на основе количественных и качественных значений атрибутов данных (полей записей). Применение кластерного анализа к объектам, представленным записями данных, позволяет автоматически разнести хранящиеся в массивах записи по различным однородным сегментам – кластерам. Однородность любого выделенного сегмента обусловлена тем, что он состоит из записей, обладающих общими свойствами, т.е. подобных записей. Группировка однородных записей в кластеры позволяет во многих случаях перейти от обработки всего массива записей к анализу небольшого числа кластеров. Таким образом, применение кластерного анализа даёт возможность резко сократить, сжать большие объёмы информации в хранилищах и базах данных и сделать эти массивы компактными и наглядными для дальнейшего использования.

Кластерный анализ применяют при решении большого числа задач в различных областях деятельности:

– в сфере маркетинга – для сегментирования рынка (выявление закономерностей в покупках, совершаемых клиентами; выделение групп потребителей со схожими стереотипами поведения и т.п.);

– в банковском деле – для определения типичных групп (профилей) добросовестных и неблагонадёжных заёмщиков;

– в страховом бизнесе – для получения профилей клиентов (с целью определения услуг страхования, обеспечивающих наименьшие для компании риски);

– в медицине – для выявления типичных клинических случаев и классификации медико-биологических объектов;

– в телекоммуникационном бизнесе – для поиска родственных групп клиентов с похожими типами пользования услугами (с целью разработки привлекательных наборов цен и услуг);

– в социологии – для обработки результатов опросов общественного мнения.

С точки зрения априорной информации о числе кластеров, на которое требуется разбить исследуемую совокупность объектов, задачи кластерного анализа можно подразделить на следующие основные типы:

– число кластеров априори задано;

– число кластеров неизвестно и подлежит определению;

– число кластеров неизвестно, но его определение не является условием решения задачи, а необходимо построить иерархическое дерево (дендрограмму) разбиения анализируемой совокупности объектов на кластеры. В данном случае требуется осуществить иерархическую кластеризацию, т.е. построить иерархическое дерево разбиения (дендрограмму) анализируемой совокупности объектов на кластеры. Дендрограммой называется такая последовательность разбиений, в которой каждое разбиение вложено в последующее разбиение в последовательности.

Результатом кластерного анализа является как выделение самих кластеров, так и определение принадлежности каждого объекта к одному из них. Часто результаты выполненного кластерного анализа являются отправной точкой для дальнейшего проведения интеллектуального анализа данных. С помощью этого дальнейшего анализа пытаются установить: что означает выявленное разбиение на кластеры и чем оно вызвано; кто является типичным «представителем» каждого кластера; с помощью каких «представителей» кластеров следует решать различные проблемные задачи и др.

2. ФОРМАЛИЗАЦИЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ

В процессе кластеризации осуществляется группировка объектов, к которым можно отнести всё, что угодно, включая наблюдения и события.

Состояние исследуемого объекта может быть описано с помощью вектора дескрипторов или многомерного набора зафиксированных на нём признаков:

$$X = \{x^1, x^2, ..., x^p\}.$$

Тогда X_i – результат измерения этих признаков на i-ом объекте. Часть признаков может носить количественный характер и принимать любые действительные значения. Другая часть носит качественный характер и позволяет упорядочивать объекты по степени проявления какого-либо качества (например, бинарный признак, отображающий присутствие или отсутствие данного свойства).

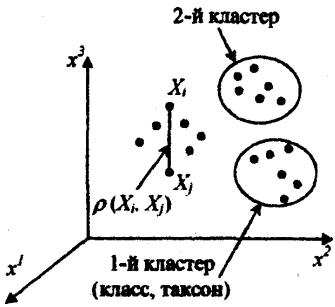

Рис. 1. Геометрическая интерпретация кластеров

Очевидно, что любое многомерное наблюдение может быть геометрически интерпретировано в виде точки в p-мерном пространстве (см. рис. 1). Естественно предположить, что геометрическая близость двух или нескольких точек в этом пространстве означает принадлежность этих точек к одному кластеру.

Чтобы решить задачу кластеризации алгоритмически, необходимо количественно определить понятие сходства и разнородности объектов. Тогда объекты X_i и X_j будем относить к одному кластеру, когда расстояние между этими объектами будет достаточно малым, и к разным – если будет достаточно большим.

Таким образом, для определения «похожести» объектов необходимо ввести меру близости или расстояния между объектами.

Неотрицательная, вещественнозначная функция $\rho(X_i, X_j)$ называется *функцией расстояния (метрикой)*, если:

1. $\rho(X_i, X_j) \geq 0$ для всех X_i и X_j;
2. $\rho(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$;
3. $\rho(X_i, X_j) = \rho(X_j, X_i)$;
4. выполняется неравенство треугольника $\rho(X_i, X_j) \leq \rho(X_i, X_k) + \rho(X_k, X_j)$, где X_i, X_j, X_k – любые 3 объекта.

Существуют различные способы вычисления расстояний. Наиболее употребительна *евклидова метрика*, которая связана с интуитивным представлением о расстоянии.

$$\rho_E(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (x_i^k - x_j^k)^2}\,.$$

Хеммингово расстояние используется как мера различия объектов, задаваемых дихотомическими (бинарными) признаками. Данная мера равна числу несовпадений значений соответствующих признаков в рассматриваемых i-ом и j-ом объектах:

$$\rho_X(X_i, X_j) = \sum_{k=1}^{p} \left| x_i^k - x_j^k \right|.$$

Существуют и другие более абстрактные меры близости. Если исследуемые признаки смешанные (количественные и качественные), то необходима нормировка по всем значениям x_i^k количественных признаков x^k:

$$\frac{x_i^k}{\max\limits_i x_i^k};\ \ i = \overline{1, n},$$

которая приводит к общей евклидовой мере близости. При разработке моделей и методов кластеризации обычно исходят из того, что объекты внутри одного кластера должны быть близки друг к другу и далеки от объектов, вошедших в другие кластеры. Точность кластеризации определяется тем, насколько близки объекты одного кластера и насколько удалены объекты, принадлежащие разным кластерам.

3. АЛГОРИТМ КЛАСТЕРИЗАЦИИ

Пусть результаты измерений n объектов представлены в виде матрицы данных размером $p \times n$, в которой множество строк представляет объекты, а множество столбцов – признаки.

$$\begin{matrix} X_1 \\ X_2 \\ \dots \\ X_n \end{matrix} \rightarrow \begin{pmatrix} x_1^1 & x_1^2 & \dots & x_1^p \\ x_2^1 & x_2^2 & \dots & x_2^p \\ \dots & \dots & \dots & \dots \\ x_n^1 & x_n^2 & \dots & x_n^p \end{pmatrix}.$$

Тогда близость между парами объектов можно представить в виде симметричной матрицы расстояний:

$$R = \begin{pmatrix} 0 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 0 & \dots & \rho_{2n} \\ \dots & \dots & \dots & \dots \\ \rho_{n1} & \rho_{n2} & \dots & 0 \end{pmatrix}.$$

В матрице R $\rho_{ii} = 0$, где $i = \overline{1,n}$.

Исследуемый в данной работе алгоритм кластеризации основан на понятии минимального остовного дерева, построенного с использованием матрицы расстояний R. В разделе 4 представлены алгоритмы Крускала и Прима построения такого дерева.

Общий алгоритм кластерного анализа, использующий подалгоритм построения минимального остовного дерева, содержит следующие основные шаги:

Шаг 0. [Инициализация] Построение матрицы расстояний (близости) R по результатам измерений n объектов, представленным матрицей данных размером $p \times n$.

Шаг 1. [Построение минимального остовного дерева] С использованием матрицы R осуществляется построение минимального остовного дерева T. Для построения минимального остовного дерева предлагается воспользоваться алгоритмами Крускала и Прима, описанными в разделе 4.

Пусть $\{d_1, d_2, \dots, d_{n-1}\}$ – множество весов (длин) рёбер минимального остовного дерева. На рис. 2 представлен пример минимального остовного дерева, построенного для 7 объектов.

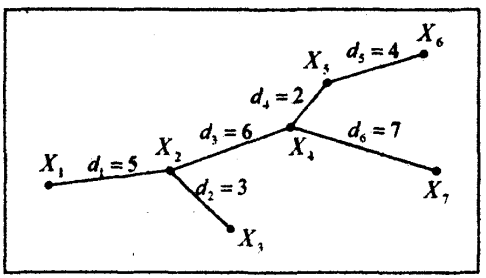

Рис. 2. Пример минимального остовного дерева T

Шаг 2. [Группировка объектов в кластеры] Вершины – объекты минимального остовного дерева группируются в кластеры.

Выбираются два объекта, которым соответствует минимальное ребро $\min_j d_j$, где $j=\overline{1,n-1}$. Далее эти объекты стягиваются в один кластер (класс, таксон, страту) и процедура шага 2 повторяется до тех пор, пока на $n-1$ этапе группирования не будет сформирован один кластер, объединяющий все объекты. STOP.

На рис. 3 представлена последовательность группировки объектов в кластеры для заданного на рис. 2 примера минимального остовного дерева. Порядок объединения объектов в кластеры отображён на рёбрах, которые связывают объединяемые объекты (см. рис. 3.1-3.7). Таким образом, первыми объединяются объекты X_4 и X_5, которые в T связывает минимальное ребро d_4 с весом 2 (см. рис. 2 и 3.1). Вторыми объединяются объекты X_2 и X_3, связанные ребром d_2 с весом 3 (см. рис. 2 и 3.2), и так далее, пока на шестом этапе группирования ранее связанные объекты (X_1, X_2, X_3, X_4, X_5, X_6) не будут объединены с объектом X_7 ребром с весом 7 (см. рис. 2 и 3.6).

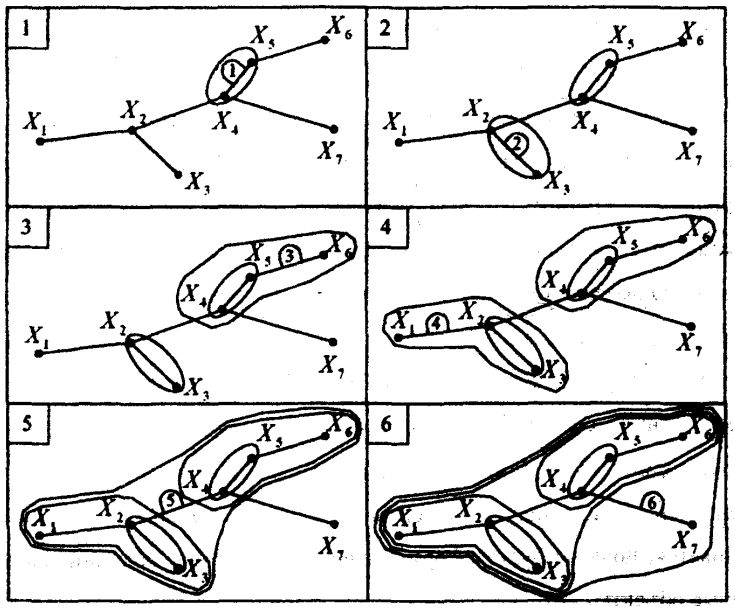

Рис. 3. Последовательность группировки объектов в кластеры

Порядок объединения объектов в кластеры может быть задан с помощью скобочного описания. Для рассматриваемого примера такая скобочная запись имеет следующий вид:

$$\Big(\big(\big((X_4, X_5), X_6\big), \big((X_2, X_3), X_1\big)\big), X_7\Big).$$

Наиболее удобным и распространённым способом описания результатов иерархической кластеризации является дендрограмма, изображённая для рассматриваемого примера на рис. 4.

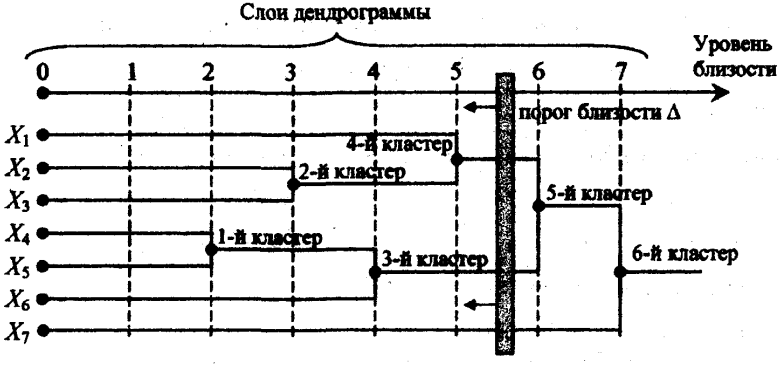

Рис. 4. Дендрограмма результатов иерархической кластеризации

Дендрограмма имеет специальную структуру дерева, состоящего из слоёв вершин, любая из которых представляет один кластер. Каждый слой вершин характеризуется своим уровнем близости. Расположение произвольной вершины – кластера относительно слоёв дендрограммы определяется её уровнем близости, который измеряется весом последнего стягиваемого ребра при образовании данного кластера.

Формирование дендрограммы начинается со слоя нулевого уровня близости, в котором каждый из исходных объектов помещается в отдельный кластер. Линии, соединяющие вершины, формируют кластеры, которые вложены один в другой. В целом, дендрограмма отражает порядок вложенности кластеров, в котором число кластеров последовательно уменьшается, пока не будет сформирован один кластер, объединяющий все исходные объекты.

Срез дендрограммы, определяемый её порогом близости Δ, используется для проведения кластерного анализа на заданное число кластеров. С этой целью порог близости Δ последовательно уменьшается от максимально возможного значения до нуля. При таком уменьшении Δ дендрограмма последовательно распадается сначала на два кластера, затем на три и т.д., пока не будут выполнены требования к необходимому числу кластеров.

4. АЛГОРИТМЫ ПОСТРОЕНИЯ МИНИМАЛЬНОГО ОСТОВНОГО ДЕРЕВА (МОД)

4.1. Общие сведения о задаче построения МОД

Действия алгоритмов построения минимального остовного дерева T рассмотрим на конкретных примерах матрицы расстояний R.

Пусть задана симметричная матрица расстояний R:

$$R = \begin{vmatrix} 0 & 11 & 9 & 7 & 8 \\ 11 & 0 & 15 & 14 & 13 \\ 9 & 15 & 0 & 12 & 14 \\ 7 & 14 & 12 & 0 & 6 \\ 8 & 13 & 14 & 6 & 0 \end{vmatrix},$$

которой можно поставить в соответствие взвешенную полносвязную сеть G с n=5 вершинами и m=10 рёбрами, представленную на рис. 5.

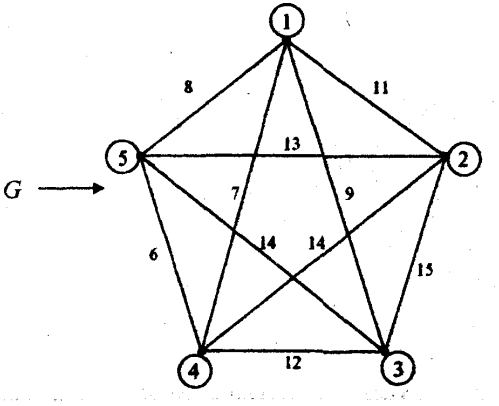

Рис. 5. Исходная сеть G для построения МОД

Тогда минимальным остовным деревом T сети G является самая дешёвая подсеть, т.е. подсеть минимального веса, которая покрывает все вершины сети G и не содержит циклов. Очевидно, что такая подсеть является деревом.

Для построения минимального остовного дерева T во взвешенной, связной и полной сети G с n вершинами и m рёбрами можно использовать ряд алгоритмов, среди которых наиболее известными являются алгоритмы Крускала и Прима.

4.2. Алгоритм Крускала

Данный алгоритм содержит следующие основные шаги:

Шаг 0. [Инициализация] Создаём сеть T с n вершинами, но без рёбер. Создаём сеть H идентичную сети G.

Шаг 1. [Цикл] До тех пор, пока сеть T не является связной сетью выполнять шаг 2, в противном случае STOP.

Шаг 2. [Отыскание ребра с наименьшим весом] Пусть (u,v) – ребро с наименьшим весом в сети H. Если при добавлении ребра (u,v) к сети T в последней не образуется циклов, то это ребро добавляется к T.

Шаг 3. [Удаление (u,v) из H] Удаляем ребро (u,v) из сети H.

На рис. 6 представлен пример построения с помощью алгоритма Крускала минимального остовного дерева T для исходной сети G, изображённой на рис. 5.

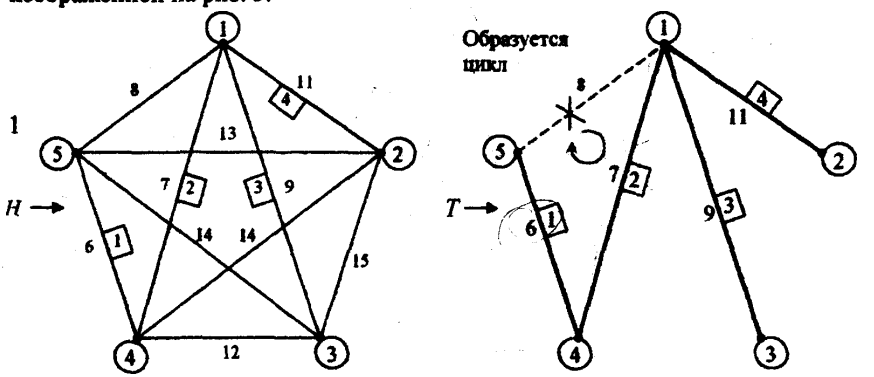

Рис.6. Пример построения минимального остовного дерева по алгоритму Крускала

Порядок присоединения рёбер (u,v) к сети T и удаления этих рёбер из сети H указан цифрами в квадратах на соответствующих рёбрах этих сетей. Из рис. 6 видно, что на третьем этапе претендентом на присоединение к T является ребро (5,1), как имеющее на данном этапе в сети H наименьший вес 8. Однако его присоединение к T не происходит, так как оно приводит к появлению цикла (1,4,5,1), и на третьем этапе присоединяется следующее по значению ребро (1,3) с весом 9.

Необходимость решения вопросов о связности сети и наличии в ней циклов делают алгоритм Крускала недостаточно эффективным. Следующий алгоритм, разработанный Примом, гарантирует построение минимального остовного дерева без проведения проверок создаваемой сети на связность и наличие в ней циклов.

4.3. Алгоритм Прима

С помощью алгоритма Прима минимальное остовное дерево порождается посредством разрастания одного поддерева от выбранной вершины. Алгоритм реализуется путём прибавления рёбер, причём добавляемое ребро должно иметь наименьший вес. Процедура выполняется на сети G с n вершинами и m рёбрами до тех пор, пока число рёбер в сети T не станет равным n-1.

Алгоритм реализует следующие основные шаги:

Шаг 0. [Инициализация] Помечаем все вершины «невыбранными». Создаём сеть T с n вершинами, но без рёбер. Выбираем произвольную вершину и помечаем её «выбранной».

Шаг 1. [Цикл] До тех пор, пока существуют «невыбранные» вершины, выполнять шаг 2, в противном случае – STOP.

Шаг 2. [Отыскание ребра с наименьшим весом] Пусть (u,v) – ребро с наименьшим весом между произвольно выбранной вершиной u и произвольной невыбранной вершиной v. Помечаем v как «выбранную» и добавляем ребро (u,v) в сеть T.

На рис. 7 представлен пример построения с помощью алгоритма Прима минимального остовного дерева T для исходной сети G, изображённой на рис. 5. Порядок добавления ребер указан цифрами в квадратах на соответствующих рёбрах порождаемой сети T. Выбранные вершины также помечены квадратами.

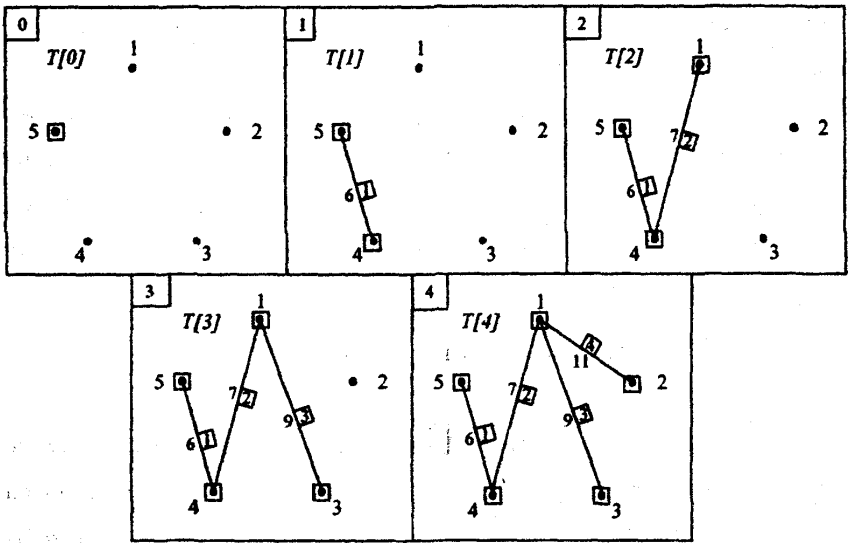

Рис. 7. Построение минимального остовного дерева с помощью алгоритма Прима

Заметим, что исходные сети G для работы алгоритмов Крускала и Прима не обязательно ограничивать только классом полных сетей. Отсутствующим рёбрам следует приписать бесконечный вес.

5. СОЗДАНИЕ ИНТЕРФЕЙСА ПРОГРАММЫ КЛАСТЕРНОГО АНАЛИЗА

При программной реализации алгоритма кластерного анализа к интерфейсу программы предъявляются следующие требования:

- возможность изменения параметров и записей предметной области;
- наглядность представления исходных данных;
- доступность результата работы программы.

В качестве программной среды удобно использовать Borland Delphi версии 6.0 и выше, так как предыдущие версии не содержат некоторых удобных функций.

Для выполнения поставленных требований рекомендуется использовать следующие компоненты среды Borland Delphi:

– SpinEdit – для изменения количества параметров и записей;

– StringGrid – для отображения в табличной форме матрицы расстояний и матрицы остовного дерева;

– PageControl и TabSheet – для отображения информации о всех признаках;

– Image – для вывода результатов иерархической кластеризации;

– TrackBar и SpinEdit – для задания порога близости;

– Memo – для вывода результатов классификации.

На рис. 8 представлен пример интерфейса программы кластерного анализа.

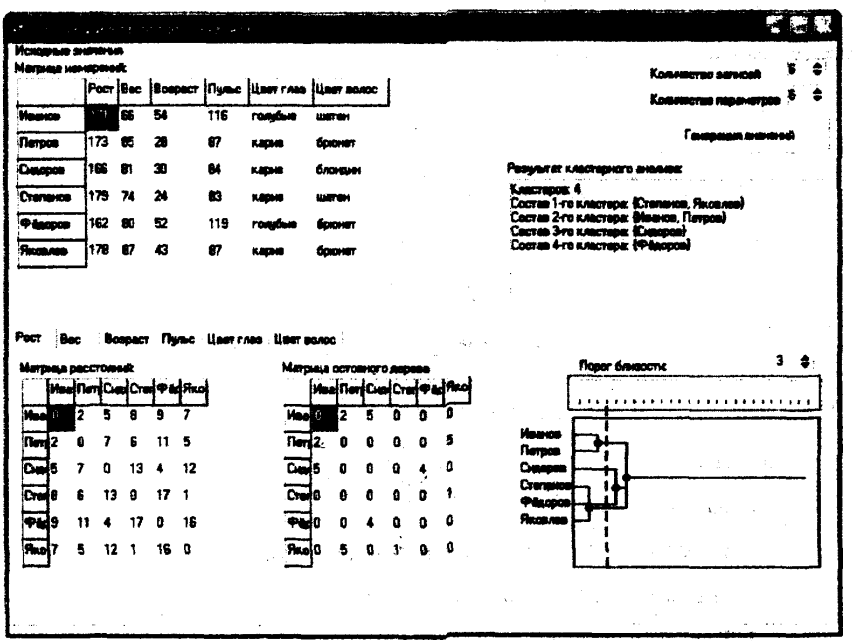

Рис.8. Пример интерфейса программы кластерного анализа

6. СОДЕРЖАНИЕ РАБОТЫ И РАБОЧЕЕ ЗАДАНИЕ

В данной лабораторной работе предлагается программно реализовать один из алгоритмов кластерного анализа при помощи среды разработки Borland Delphi, а также, в соответствии с заданными вариантами предметной области, осуществить компьютерное проведение самого кластерного анализа.

В ходе работы необходимо выполнить следующее рабочее задание:

1. Изучить различные виды алгоритмов кластерного анализа, отличающиеся подалгоритмами построения минимального остовного дерева.

2. Изучить заданный вариант предметной области кластеризации, представленный в виде таблиц данных.

3. Выполнить следующие этапы программной реализации алгоритма кластерного анализа:

– составить алгоритм работы программы;

– создать интерфейс программы, позволяющий реализовать кластерный анализ (см. раздел 5);

– в соответствии с заданным вариантом предметной области программно реализовать один из алгоритмов кластерного анализа;

– ввести исходные данные предметной области в программу;

– получить результаты компьютерного проведения кластерного анализа при разных значениях исходных данных.

7. ВАРИАНТЫ ЗАДАНИЙ

7.1. Вариант 1. Сортировочная станция.

Признаки к варианту 1:

1. Вагонооборот (ваг/год);
2. Затраты на переработку и накопление (вагоно-часы/год);
3. Транзит без переработки (ваг/год);
4. Транзит с переработкой (ваг/год);
5. Количество сортировочных путей.

Таблица 1. Условные значения признаков к варианту 1

Наименование станции	Вагонооборот (тыс. ваг/год)	Затраты на накопление и переработку (тыс.вагоно-часов/год)	Транзит без переработки (тыс. ваг/год)	Транзит с переработкой (тыс. ваг/год)	Количество сортировочных путей
Лянгасово	49850	3450	1965	2540	20
Егоршино	42361	2985	1788	2394	18
Орехово	13428	905	2136	1521	14
Пермь	37694	3054	1323	1447	15
Войновка	22748	1985	984	840	17
Инская	20016	2130	1980	1922	14
Омск	44689	3169	1562	1065	17
Екатеринбург	39460	4126	650	2008	14
Бердяуш	23164	2674	2654	651	9
Смычка	10748	798	846	2013	7
Чусовская	29710	3012	1654	654	10
Агрыз	11547	1065	764	1024	11
Нижний Новгород	38155	3316	1545	1985	15
Юдино	17568	897	862	331	5
Челябинск	40036	3155	2560	606	14
Богданович	23757	2581	1035	1654	9
Каменск-Уральский	14692	964	1876	1065	4
Курган	19864	2379	1654	1689	8
Алтайская	13000	988	2985	1986	12

7.2. Вариант 2. Станционный перегон.

Признаки к варианту 2:

1. Длина (км);
2. Время с последнего капитального ремонта (дней);
3. Электрификация перегона;
4. Пропускная способность (ваг/год);
5. Количество путей.

Таблица 2. Условные значения признаков к варианту 2

Наименование перегона	Длина (км)	Время с последнего капитального ремонта (дней)	Электрификация перегона	Пропускная способность (ваг/год)	Количество путей
Березники-Чусовская	376	904	Да	1985	3
Смычка-Егоршино	285	1032	Да	655	2
Омск-Войновка	136	844	Нет	3204	2
Алтайская-Курган	582	642	Нет	654	1
Омск-Курган	42	165	Нет	4510	2
Лянгасово-Нижний Новгород	390	640	Да	1645	1

Таблица 2. (Продолжение)

Санкт-Петербург-Лянгасово	489	564	Нет	5413	2
Лянгасово-Пермь	376	846	Нет	584	2
Пермь-Агрыз	109	1002	Да	1534	3
Агрыз-Екатеринбург	110	455	Да	3520	2
Смычка-Гороблагодатская	143	1364	Нет	1104	1
Войновка-Богданович	122	99	Нет	1246	2
Орехово-Юдино	316	961	Да	854	1
Омск-Алтайская	262	156	Да	3510	3
Богданович-Каменск-Уральский	328	246	Нет	946	3
Серов-Егоршино	561	1062	Да	1581	2
Челябинск-Каменск-Уральский	109	346	Да	796	3
Серов-Гороблагодатская	161	964	Нет	779	1

7.3. Вариант 3. Вагонный парк.

Признаки к варианту 3:

1. Пробег вагона (км);
2. Грузовладелец (ОАО «РЖД», физическое или юридическое лицо);
3. Время после ремонта (дней);
4. Тоннаж (т);
5. Вид вагона (пассажирский, полувагон, цистерна, рефрижератор).

Таблица 3. Условные значения признаков к варианту 3

Номер вагона	Пробег вагона (км)	Грузовладелец	Время после ремонта (дней)	Тоннаж (т)	Вид вагона
2210511	32100	ООО «АЗОТ»	1620	50	Пассажирский
8005924	5200	ОАО «РЖД»	206	34	Рефрижератор
8011456	16958	ОАО «РЖД»	6589	34	рефрижератор
7940061	20695	Иванов А.А.	10264	50	цистерна
6980469	36599	ОАО «РЖД»	11336	46	полувагон

7.4. Вариант 4. Студенческий состав.

Признаки к варианту 4:

1. Рост (см);
2. Вес (кг);
3. Возраст (лет);

4. Уровень интеллекта (IQ);

5. Образование (среднее, высшее, учёная степень).

Таблица 4. Условные значения признаков к варианту 4

Фамилия	Рост (см)	Вес (кг)	Возраст (лет)	Уровень интеллекта	Образование
Иванов	165	65	18	70	Среднее
Петров	182	112	65	142	Уч. степень
Сидоров	169	95	27	100	Высшее
Степанов	176	74	32	94	Высшее
Фёдоров	189	82	40	82	Среднее

7.5. Вариант 5. Кредитная информация.

Признаки к варианту 5:

1. Возраст (лет);

2. Заработок (руб);

3. Кредитная история;

4. Семейное положение;

5. Образование (среднее, высшее, учёная степень).

Таблица 5. Условные значения признаков к варианту 5

Фамилия	Возраст (лет)	Заработок (руб)	Кредитная история	Семейное положение	Образование
Иванов	23	15000	Нет	Женат	Высшее
Петров	50	40000	Есть	Женат	Высшее
Сидоров	34	16000	Есть	Холост	Среднее
Фёдоров	29	20000	Нет	Холост	Среднее
Яковлев	42	18000	Есть	Холост	Уч. степень

8. СОДЕРЖАНИЕ ОТЧЁТА

Отчёт должен содержать:

1. Название, цель работы, вариант задания.

2. Листинг программы кластерного анализа.

3. Результаты работы программы (скриншоты).

9. КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В чём состоит задача кластерного анализа?
2. К какой предметной области может быть применён кластерный анализ?
3. В чём заключается практическая ценность кластерного анализа?
4. Как классифицируются задачи кластерного анализа с точки зрения информации о числе кластеров?
5. Что является результатом кластерного анализа?
6. Как количественно измеряется «похожесть» объектов при кластерном анализе?
7. Как осуществляется группировка объектов в кластеры с помощью исследуемого алгоритма?
8. Какая структура является наиболее удобным и распространённым способом описания результатов иерархической кластеризации?
9. В чём состоит задача построения минимального остовного дерева?
10. Какие требования предъявляются к интерфейсу программы кластерного анализа?

10. ЛИТЕРАТУРА

1. Кутыркин А.В. Модели и методы разработки крупномасштабных предметных областей управления транспортными системами и производством: Монография. – М.: МИИТ, 2004. – 148 с.
2. Жамбю М. Иерархический кластер-анализ и соответствия. – М.: Финансы и статистика, 1998. – 342 с.
3. Смирнов Е.С. Таксономический анализ. – М.: МГУ, 1969. – 352 с.