**Лабораторная работа № 8**

**Кластерный анализ в системе STATISTICA**

В программе STATISTICA доступны следующие меры сходства объектов: евклидова метрика, квадрат евклидовой метрики, манхэттенское расстояние, или «расстояние городских кварталов», метрика Чебышева, метрика Минковского, пирсоновский коэффициент корреляции (точнее, 1–пирсоновский коэффициент корреляции), коэффициент совстречаемости (точнее, 1–коэффициент совстречаемости).

В STATISTICA реализованы следующие методы кластеризации – агломеративные методы: **joining (tree clustering)**, **two-way joining**, а также метод **k-средних (k-means clustering)**.

Обычно перед началом классификации данные стандартизуются (вычитается среднее и производится деление на корень квадратный из дисперсии). Полученные в результате стандартизации переменные имеют нулевое среднее и единичную дисперсию. Рассматриваемые далее данные уже стандартизованы.

В STATISTICA можно выбрать следующие правила иерархического объединения кластеров:

− **Single linkage** – метод одиночной связи;

− **Complete linkage** – метод полной связи;

− **Unweighted pair group average** – невзвешенный метод «средней связи»;

− **Weighted pair group average** – взвешенный метод «средней связи»;

− **Weighted centroid pair group (median)** – взвешенный центроидный метод;

− **Ward method** – метод Уорда.

|  |
| --- |
|  |
| *Рис. 1 – Запуск модуля Кластерный анализ* |

В примере рассматриваются автомобили разных марок, которые различаются ценой, рас-ходом горючего и неко-торыми техническими характеристиками.

В рабочем окне STATISTICA выберете название модуля – **Cluster Analysis (кластерный анализ)**, высветите его имя и щелкните на его имени (рис. 1).

На экране появится стартовая панель модуля **Claster Analysis (кластерный анализ)** (рис. 2).



*Рис. 2 – Стартовая панель модуля Кластерный анализ*

Стандартным образом, нажав **кнопку Open Data**, откройте окно выбора файла (рисунок 3).



*Рис. 3 – Выбор файла с данными об автомобилях*

Выберите в этом окне файл **Cars.sta** и дважды нажмите левую кнопку мыши. Файл выбран, и вы вернетесь обратно, в стартовую панель модуля.

В рабочем окне, сзади стартовой панели, вы увидите открытый файл с данными (рисунок 4).



*Рис. 4 – Файл cars.sta с данными автомобилей разных марок*

Из информации в верхней части окна видно, что в файле **Cars.sta** записаны цена автомобиля, технические характеристики, количество миль, пройденных на одном галлоне бензина.

Всего в файле содержатся данные о 22 машинах разных марок. Марки машин – это случаи.

Переменные в этом файле:

- **PRICE** – цена;

- **ACCELE** - **HANDLI** – технические характеристики;

- **MILAGE** – расход горючего (количество миль, пройденных на одном галлоне бензина).

Все характеристики машин уже стандартизованы (из значений переменной price вычтена средняя цена и разность поделена на корень квадратный из дисперсии).

Задача состоит в том, чтобы разбить автомобили на несколько групп, в которых автомобили мало отличаются друг от друга (существенно меньше, чем в целом в совокупности).

Задача эта сложна, так как сравниваются машины не по какому-то одному параметру, а по нескольким параметрам одновременно. По одним характеристикам одни машины близки друг к другу, по другим – другие. В конечном итоге разбиение на группы – тоже не самоцель. Конечно, число параметров можно увеличить. Очевидно, разбив машины на группы, можно лучше в целом представить их совокупность, с тем, чтобы затем более обоснованно принимать решение, например при покупке или обмене одной машины на другую.

Если бы машины сравнивались по одному параметру, например по расходу горючего, то, наверное, следовало бы выбрать машину с меньшим расходом топлива на одну милю. Все машины были бы упорядочены в одну линию, и задача не представляла бы проблем.

Однако параметров несколько, и ситуация существенно усложняется.

**Выбор метода**

Посмотрите на стартовую панель. В главной ее части находится список методов кластерного анализа, реализованных в STATISTICA. В списке методов высветите **k-means (k-средних)** (рис. 2) и нажмите кнопку **OK** в правом верхнем углу панели. Диалоговое окно метода **k-means** появится на экране (рисунок 5).



*Рис. 5 – Диалоговое окно метода k-means*

Начните работать в данном окне. Прежде всего, выберите переменные для анализа. Нажмите кнопку **Variables (переменные)** в левом верхнем углу текущего окна и откройте диалоговое окно **Select variable for the analysis (выбрать переменные для анализа)** (рис. 6).



*Рис. 6 – Выбор переменных для кластерного анализа*

Так как машины разбиты на группы и учитываются все параметры, то нажмите вначале кнопку **Select All (выбрать все)**, а затем нажмите кнопку **OK**. Посмотрите далее на поле **Cluster (кластер)**, находящееся ниже кнопки **Variables (переменные)** (рис. 5). Нажав на стрелку в этом поле, выберите пункт меню **Cases (случаи)**. Альтернативный выбор был бы **Variables (переменные)**. Так следует поступить, если нужно кластеризировать переменные.

В данном примере кластеризируются машины, которые являются случаями в исходном файле данных, поэтому выбирается пункт **Cases**.

В поле **Number of clusters (число кластеров)** нужно определить число групп, на которые необходимо разбить автомобили. Запишите в это поле число 3 (таким образом, машины разбиваются на 3 кластера).

В строке **Number of iterations (Число итераций)** задается максимальное число итераций, используемых при построении классов. Задайте, например, число 10.

В строке **Missing data** задается способ обработки пропущенных значений в данных (например, для какой-то машины отсутствует значение некоторого параметра). В данном примере пропусков в данных нет и обработки пропущенных значений не происходит.

Группа опций **Initial cluster centers** позволяет задать начальные центры кластеров.

Сделайте установки, как показано на рис. 5. После того как все установки сделаны, нажмите кнопку **ОК** в верхнем правом углу окна **k-means Clustering** и запустите вычислительную процедуру. Спустя несколько секунд после нажатия кнопки **ОК** в **k-means Clustering** окно результатов появится на экране (рис. 7).



*Рис. 7 – Окно результатов кластеризации машин по методу средних*

В верхней части окна записана информация: число переменных, число случаев, метод кластеризации, число кластеров, а также сообщение о том, после скольких итераций найдено решение: **Solution was obtained after 2 iterations (решение найдено после 2 итераций)**.

Кнопки в нижней части окна позволяют провести анализ результатов кластеризации.

Кнопка **Analysis of variation (дисперсионный анализ)** позволяет просмотреть таблицу дисперсионного анализа.

Кнопка **Cluster Means&Euclidean Distances** позволяет вывести таблицы, в первой из которых указаны средние для каждого кластера (усреднение производится внутри кластера), во второй указаны евклидовы расстояния и квадраты евклидовых расстояний между кластерами.

Кнопка **Graph of means** позволяет посмотреть средние значения для каждого кластера на линейном графике.

Кнопка **Descriptive Statistics for each clusters** открывает электронную таблицу с описательными статистиками для каждого кластера (среднее, дисперсия и т. д.)

Кнопка **Save classifications and distances** позволяет сохранить результаты классификации в файле STATISTICA для дальнейшего исследования.

Следует посмотреть, как распределились машины по кластерам. Нажмите для этого кнопку **Member of each cluster&distances**. На экране появятся 3 электронные таблицы с названиями машин, отнесенных к определенным кластерам (рис. 8–10).



*Рис. 8 – Первый кластер*



*Рис. 9 – Второй кластер*



*Рис. 10 – Третий кластер*

В строках таблиц указано расстояние от каждой машины до центра кластера.

Нажмите на кнопку **Cluster means&Euclidean distances**. На экране появится таблица, в которой даны евклидовы расстояния между средними кластеров (по каждому из параметров внутри кластера вычисляется среднее, получается 3 точки в пятимерном пространстве, и между ними находится расстояние) (рис. 11).



*Рис. 11 – Расстояния между кластерами*

Из таблицы видно, что расстояние между кластерами даны под диагональю – между 1 и 2 кластером 0,969, а например, между вторым и третьим – 1,876. Над диагональю в таблице даны квадраты расстояний между кластерами.

С помощью кнопки **Graph of means (график средних)** строятся следующие графики средних значений характеристик машин для каждого кластера (рис. 12).

В системе реализованы также и другие методы кластеризации, в частности так называемый **two-way joining**, в котором кластеризируются случаи и переменные одновременно. Если вы воспользуетесь **Joining (tree clustering)**, то сможете увидеть дендрограмму, или дерево объединения (рис. 13).

|  |  |
| --- | --- |
|  |  |
| *Рис. 12 – График средних для каждого кластера* | *Рис. 13 – Дерево объединения машин разных марок в кластер методом одиночной связи* |

**Задание.** Используя данные социально-экономического развития и загрязнённости радионуклидами (средние значения суммарной годовой эффективной индивидуальной дозы и удельной активности молока) 46 населённых пунктов, классифицируйте их при помощи кластерного анализа на 3 группы и дайте рекомендации о дальнейшей судьбе этих населённых пунктов.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **№ НП** | **Соц.-экон.** | **Радиолог.** | **№ НП** | **Соц.-экон.** | **Радиолог.** | **№ НП** | **Соц.-экон.** | **Радиолог.** |
| 1 | 0,247193289 | 0,295400607 | 17 | 0,715452937 | 0,354404142 | 33 | 0,8274442 | 0,361130457 |
| 2 | 0,49097333 | 0,495497008 | 18 | 0,885349202 | 0,572964656 | 34 | 0,894219018 | 1,000000000 |
| 3 | 0,768153128 | 0,240149383 | 19 | 0,658321153 | 0,277728731 | 35 | 0,38097797 | 0,458738374 |
| 4 | 0,83789641 | 0,354305136 | 20 | 0,738969511 | 0,632279179 | 36 | 0,7502056 | 0,30138691 |
| 5 | 0,920873425 | 0,319777147 | 21 | 0,683658507 | 0,357111189 | 37 | 0,613001789 | 0,512266066 |
| 6 | 0,836931986 | 0,408670359 | 22 | 0,683658507 | 0,591685683 | 38 | 0,442903759 | 0,369355736 |
| 7 | 0,642086133 | 0,30919644 | 23 | 0,511566822 | 0,428064385 | 39 | 0,852623864 | 0,230988926 |
| 8 | 0,75447239 | 0,287696776 | 24 | 0,763286867 | 0,77716795 | 40 | 0,789770272 | 0,397828781 |
| 9 | 0,844316587 | 0,339890945 | 25 | 0,451861817 | 0,339317409 | 41 | 0,74822401 | 0,526758725 |
| 10 | 0,43312923 | 0,242211888 | 26 | 0,46782343 | 0,428489112 | 42 | 0,892492356 | 0,369276556 |
| 11 | 0,922549745 | 0,289111366 | 27 | 0,849075263 | 0,298006857 | 43 | 0,737587506 | 0,657796926 |
| 12 | 0,823975995 | 0,242543293 | 28 | 0,760966155 | 0,251519518 | 44 | 0,612246087 | 0,242905587 |
| 13 | 0,962962191 | 0,334658522 | 29 | 0,298413449 | 0,201345166 | 45 | 0,773756059 | 0,23284717 |
| 14 | 0,800143159 | 0,222998021 | 30 | 0,820472659 | 0,438164981 | 46 | 0,932885374 | 0,270838401 |
| 15 | 0,828420843 | 0,695622825 | 31 | 0,886396124 | 0,251515902 |  |  |  |
| 16 | 1,000000000 | 0,365521838 | 32 | 0,760527189 | 0,533589347 |  |  |  |