

Лабораторная работа #2

Машинное обучение. Кластерный анализ

Цель: изучить основы методов Machine Learning в контексте задачи кластерного анализа (cluster analysis), приобрести навыки работы с методами Machine Learning в системе STATISTICA StatSoft, осуществить обработку методами Machine Learning индивидуального набора данных и интерпретацию результатов

1 Ход работы

- 1) изучить теоретические сведения
- 2) приобрести навыки работы с методами Machine Learning в контексте задачи кластерного анализа в системе STATISTICA StatSoft, реализуя приведенный ниже пример
- 3) на основе приобретенных практических навыков осуществить все этапы обработки методами Machine Learning в контексте задачи кластерного анализа и интерпретацию результатов согласно варианту индивидуального задания
- 4) оформить отчет и подготовиться к защите лабораторной работы по полученным результатам и контрольным вопросам

2 Содержание отчета и требования к его оформлению

- 1) отчет оформляется в печатном виде
- 2) отчет содержит титульный лист, исходные данные, результаты выполнения этапов обработки данных в виде скриншотов и обязательных комментариев по ходу выполнения работы, выводы
- 3) к отчету прилагается файл исходных данных *.sta и файл проекта в электронном виде с целью осуществления выборочного контроля

3 Варианты исходных данных

- исходные данные - в файле Country-data for vars.xls (описание см. Country-data_descript.txt).

4 Краткие теоретические сведения

Кластерный анализ (cluster analysis) – совокупность многомерных статистических методов классификации объектов по характеризующим их признакам, разделение совокупности объектов на однородные группы, близкие по определяющим критериям, выделение объектов определенной группы.

Кластер – это группы объектов, выделенные в результате кластерного анализа на основе заданной меры сходства или различий между объектами. Объект – это конкретные предметы исследования, которые необходимо классифицировать. Объектами при классификации выступают, как правило, наблюдения. Например, потребители продукции, страны или регионы, товары и т.п. Хотя можно проводить кластерный анализ и по переменным. Классификация объектов в многомерном кластерном анализе происходит по нескольким признакам одновременно. Это могут быть как количественные, так и категориальные переменные в зависимости от метода кластерного анализа. Итак, главная цель кластерного анализа – нахождение групп схожих объектов в выборке.

Сфера использования кластерного анализа, из-за его универсальности, очень широка. Кластерный анализ применяют в экономике, маркетинге, археологии, медицине, психологии, химии, биологии, государственном управлении, филологии, антропологии, социологии и других областях.

Вот несколько примеров применения кластерного анализа:

медицина – классификация заболеваний, их симптомов, способов лечения, классификация групп пациентов;

маркетинг – задачи оптимизации ассортиментной линейки компании, сегментация рынка по группам товаров или потребителей, определение потенциального потребителя;

социология – разбиение респондентов на однородные группы;

психиатрия – корректная диагностика групп симптомов является решающей для успешной терапии;

биология – кластеризация организмов по группе;

экономика – кластеризация субъектов инвестиционной привлекательности.

Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве X , разбить множество объектов G на m (m – целое) кластеров (подмножеств) Q_1, Q_2, \dots, Q_m , так, чтобы каждый объект G_j принадлежал одному и только одному подмножеству разбиения. При этом объекты, принадлежащие одному и тому же кластеру, были сходными, в то время как объекты, принадлежащие разным кластерам, были разнородными.

Решением задачи кластерного анализа являются разбиения, удовлетворяющие некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок, который называют целевой функцией.

Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонения:

$$W = \sigma_n = \sum_{j=1}^n (x_j - \bar{x})^2 = \sum_{j=1}^n x_j^2 - \frac{1}{n} \left(\sum_{j=1}^n x_j \right)^2,$$

где x_j - представляет собой измерения j -го объекта.

Кластер имеет следующие *математические характеристики*: *центр, радиус, среднеквадратическое отклонение, размер кластера*.

Центр кластера - это среднее геометрическое место точек в пространстве переменных.

Радиус кластера - максимальное расстояние точек от *центра кластера*.

Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается перекрытие кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют *спорными*.

Спорный объект - это объект, который по мере сходства может быть отнесен к нескольким кластерам.

Размер кластера может быть определен либо по *радиусу кластера*, либо по *среднеквадратичному отклонению* объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до *центра кластера* меньше *радиуса кластера*. Если это условие выполняется для двух и более кластеров, объект является *спорным*.

Неоднозначность данной задачи может быть устранена экспертом или аналитиком.

Работа кластерного анализа опирается на два предположения. Первое предположение - рассматриваемые признаки объекта в принципе допускают желательное разбиение пула (совокупности) объектов на кластеры. Второе предположение - правильность выбора масштаба или единиц измерения признаков. Выбор масштаба в кластерном анализе имеет большое значение.

Эта проблема решается при помощи предварительной *стандартизации* переменных. *Стандартизация* (standardization) или нормирование (normalization) приводит значения всех преобразованных переменных к единому диапазону значений путем выражения через отношение этих значений к некой величине, отражающей определенные свойства конкретного признака.

Наряду со *стандартизацией* переменных, существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы отражал значимость соответствующей переменной. В качестве весов могут выступать экспертные оценки, полученные в ходе опроса экспертов - специалистов предметной области. Полученные произведения нормированных переменных на соответствующие веса позволяют получать расстояния между точками в многомерном пространстве с учетом неодинакового веса переменных.

В кластерном анализе для количественной оценки сходства вводится понятие *метрики*. Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается k признаками, то он может быть представлен как точка в k -мерном пространстве, и сходство с другими объектами будет определяться как соответствующее расстояние.

| Показатели | Формулы |
|-----------------------------------------------------------|-------------------------------------------------------------------------|
| Для количественных шкал | |
| Линейное расстояние | $d_{lij} = \sum_{l=1}^m x_i^l - x_j^l $ |
| Евклидово расстояние | $d_{Eij} = \left(\sum_{l=1}^m (x_i^l - x_j^l)^2 \right)^{\frac{1}{2}}$ |
| Квадрат евклидова расстояния | $d^2_{Eij} = \sum_{l=1}^m (x_i^l - x_j^l)^2$ |
| Обобщенное степенное расстояние Минковского | $d_{Pij} = \left(\sum_{l=1}^m (x_i^l - x_j^l)^P \right)^{\frac{1}{P}}$ |
| Расстояние Чебышева | $d_{ij} = \max_{1 \leq i, j \leq l} x_i - x_j $ |
| Расстояние городских кварталов (Манхэттенское расстояние) | $d_H(x_i, x_j) = \sum_{l=1}^k x_i^l - x_j^l $ |

Евклидово расстояние является самой популярной метрикой в кластерном анализе. Оно попросту является геометрическим расстоянием в многомерном пространстве. Геометрически оно лучше всего объединяет объекты в шарообразных скоплениях.

Квадрат евклидова расстояния. Для придания больших весов более отдаленным друг от друга объектам можем воспользоваться квадратом *евклидова расстояния* путем возведения в квадрат стандартного *евклидова расстояния*.

Обобщенное степенное расстояние представляет только математический интерес как универсальная метрика.

Расстояние Чебышева. Это расстояние стоит использовать, когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению.

Манхэттенское расстояние (расстояние городских кварталов), также называемое "хэмминговым" или "сити-блок" расстоянием. Это расстояние рассчитывается как среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к результатам, подобным расчетам расстояния евклида. Однако, для этой меры влияние отдельных выбросов меньше, чем при использовании *евклидова расстояния*, поскольку здесь координаты не возводятся в квадрат.

Процент несогласия. Это расстояние вычисляется, если данные являются категориальными.

Методы кластерного анализа

Методы кластерного анализа можно разделить на две группы: ·

иерархические; ·

неиерархические.

Каждая из групп включает множество подходов и алгоритмов. Используя различные методы кластерного анализа, аналитик может получить различные решения для одних и тех же данных. Это считается нормальным явлением.

Иерархические алгоритмы

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Иерархические агломеративные методы (Agglomerative Nesting, AGNES) Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров. В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (делимые) методы (DIvisive ANALysis, DIANA) Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Принцип работы описанных выше групп методов в виде *дендрограммы* показан на рис.

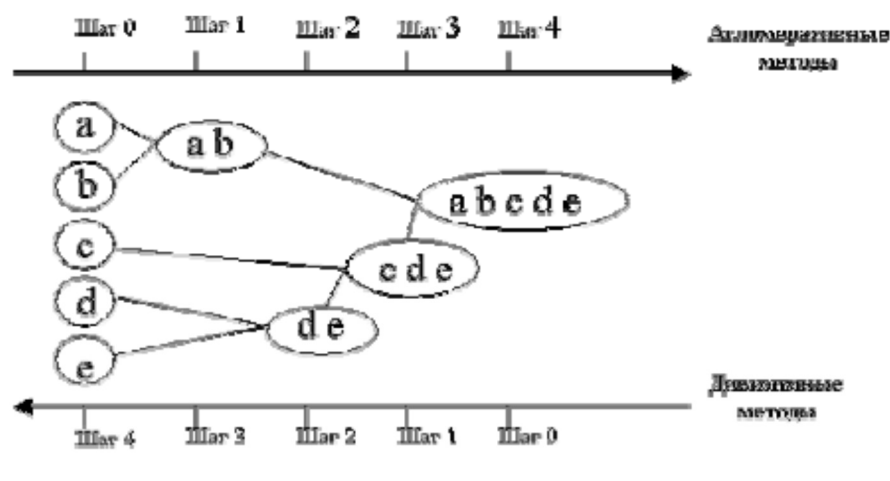
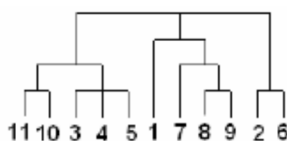


Рис. 1.2. Дендрограмма агломеративных и дивизимных методов

Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо деления на группы (дивизимные методы). Иерархические методы кластерного анализа

используются при небольших объемах наборов данных. Преимуществом иерархических методов кластеризации является их наглядность. Иерархические алгоритмы связаны с построением *дендрограмм* (от греческого dendron - "дерево"), которые являются результатом иерархического кластерного анализа. Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров. Дендрограмма (dendrogram) - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров. Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры. Дендрограмма представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии. Существует много способов построения дендограмм. В дендограмме объекты могут располагаться вертикально или горизонтально.

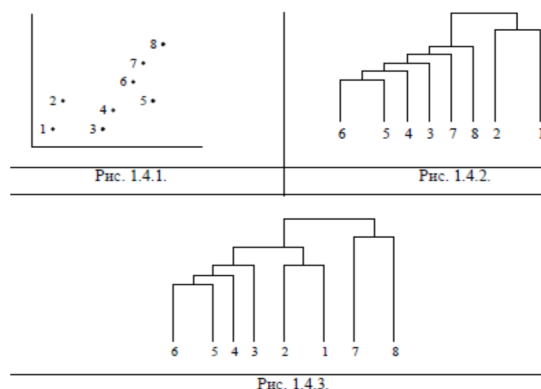


Пример вертикальной дендрограммы приведен на рис. Числа 11, 10, 3 и т.д. соответствуют номерам объектов или наблюдений исходной выборки. Мы видим, что на первом шаге каждое наблюдение представляет один кластер (вертикальная линия), на втором шаге наблюдаем объединение таких наблюдений: 11 и 10; 3, 4 и 5; 8 и 9; 2 и 6. На втором шаге продолжается объединение в кластеры: наблюдения 11, 10, 3, 4, 5 и 7, 8, 9. Данный процесс продолжается до тех пор, пока все наблюдения не объединятся в один кластер.

Обобщенная алгомеративная процедура. На первом шаге каждый объект считается отдельным кластером. На следующем шаге объединяются два ближайших объекта, которые образуют новый класс, определяются расстояния от этого класса до всех остальных объектов, и размерность матрицы расстояний D сокращается на единицу. На p -ом шаге повторяется та же процедура на матрице $D(n-p)(n-p)$, пока все объекты не объединятся в один класс.

Если сразу несколько объектов (классов) имеют минимальное расстояние, то возможны две стратегии: выбрать одну случайную пару или объединить сразу все пары. Первый способ является классическим и реализован во всех процедурах (иногда его называют восходящей иерархической классификацией). Вторым способом называют методом ближайших соседей (не путать с алг. "Ближайшего соседа") и используют реже.

Результаты работы всех иерархических процедур обычно оформляют в виде так называемой *дендограммы* (рис. 1.4.1 – 1.4.3). В дендограмме номера объектов располагаются по горизонтали, а по вертикали - результаты кластеризации.



Расстояния между кластерами

На первом шаге, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Однако, когда связываются вместе несколько объектов, возникает вопрос, как следует определить расстояния между кластерами? Другими словами, необходимо правило объединения или связи для двух кластеров. Кластерный анализ предлагает широкий выбор таких методов.

1. *Расстояние “Ближайшего соседа” (Одиночная связь).* Расстояние равно расстоянию между ближайшими объектами классов.

2. *Расстояние “Дальнего соседа” (Полная связь).* Расстояние равно расстоянию между самыми дальними объектами классов.

3. *Невзвешенное попарное среднее.* В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Метод эффективен, когда объекты в действительности формируют различные роши, однако он работает одинаково хорошо и в случаях протяженных (цепочного типа) кластеров.

4. *Взвешенное попарное среднее.* Метод идентичен методу невзвешенного попарного среднего, за исключением того, что при вычислениях размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента. Поэтому предлагаемый метод должен быть использован (скорее даже, чем предыдущий), когда предполагаются неравные размеры кластеров.

5. *Невзвешенный центроидный метод.* В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

6. *Взвешенный центроидный метод (медиана).* Тот метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них). Поэтому, если имеются (или подозреваются) значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее предыдущего.

7. *Метод Варда (Ward, 1963).* В этом методе в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений, которая есть ни что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров.

Процедуры эталонного типа

Наряду с иерархическими методами классификации, существует многочисленная группа так называемых итеративных методов кластерного анализа (метод k - средних.). Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т.д.). Название метода было предложено Дж. Мак-Куином в 1967 г. В отличие от иерархических процедур метод k - средних не требует вычисления и хранения матрицы расстояний или сходств между объектами. Алгоритм этого метода предполагает использование только исходных значений переменных. Для начала процедуры классификации должны быть заданы k выбранных объектов, которые будут служить эталонами, т.е. центрами кластеров. Считается, что алгоритмы эталонного типа удобные и быстродействующие. В этом случае важную роль играет выбор начальных условий,

которые влияют на длительность процесса классификации и на его результаты. Метод k - средних удобен для обработки больших статистических совокупностей.

Математическое описание алгоритма метода k - средних

Пусть имеется n наблюдений, каждое из которых характеризуется m признаками X_1, X_2, \dots, X_m . Эти наблюдения необходимо разбить на k кластеров.

Для начала из n точек исследуемой совокупности отбираются случайным образом или задаются исследователем исходя из каких-либо априорных соображений k точек (объектов). Эти точки принимаются за эталоны. Каждому эталону присваивается порядковый номер, который одновременно является и номером кластера.

На первом шаге из оставшихся $(n - k)$ объектов извлекается точка X_i с координатами $(x_{i1}, x_{i2}, \dots, x_{im})$ и проверяется, к какому из эталонов (центров) она находится ближе всего. Для этого используется одна из метрик, например, евклидово расстояние. Проверяемый объект присоединяется к тому центру (эталону), которому соответствует минимальное из расстояний.

Эталон заменяется новым, пересчитанным с учетом присоединенной точки, и вес его (количество объектов, входящих в данный кластер) увеличивается на единицу.

Если встречаются два или более минимальных расстояния, то i -ый объект присоединяют к центру с наименьшим порядковым номером.

На следующем шаге выбираем точку X_{i+1} и для нее повторяются все процедуры. Таким образом, через $(n - k)$ шагов все точки (объекты) совокупности окажутся отнесенными к одному из k кластеров, но на этом процесс разбиения не заканчивается.

Для того чтобы добиться устойчивости разбиения по тому же правилу, все точки X_1, X_2, \dots, X_n опять подсоединяются к полученным кластерам, при этом веса продолжают накапливаться. Новое разбиение сравнивается с предыдущим. Если они совпадают, то работа алгоритма завершается. В противном случае цикл повторяется.

Окончательное разбиение имеет центры тяжести, которые не совпадают с эталонами, их можно обозначить C_1, C_2, \dots, C_k . При этом каждая точка X_i ($i = 1, 2, \dots, n$) будет относиться к такому кластеру (классу) l , для которого расстояние минимально.

Возможны две модификации метода k - средних. Первая предполагает пересчет центра тяжести кластера после каждого изменения его состава, а вторая – лишь после того, как будет завершен просмотр всех данных. В обоих случаях итеративный алгоритм этого метода минимизирует дисперсию внутри каждого кластера, хотя в явном виде такой критерий оптимизации не используется.

5. Пример использования кластерного анализа *STATISTICA* в автостраховании

В *STATISTICA* реализованы классические методы кластерного анализа, включая методы k-средних, иерархической кластеризации и двухвходового объединения.

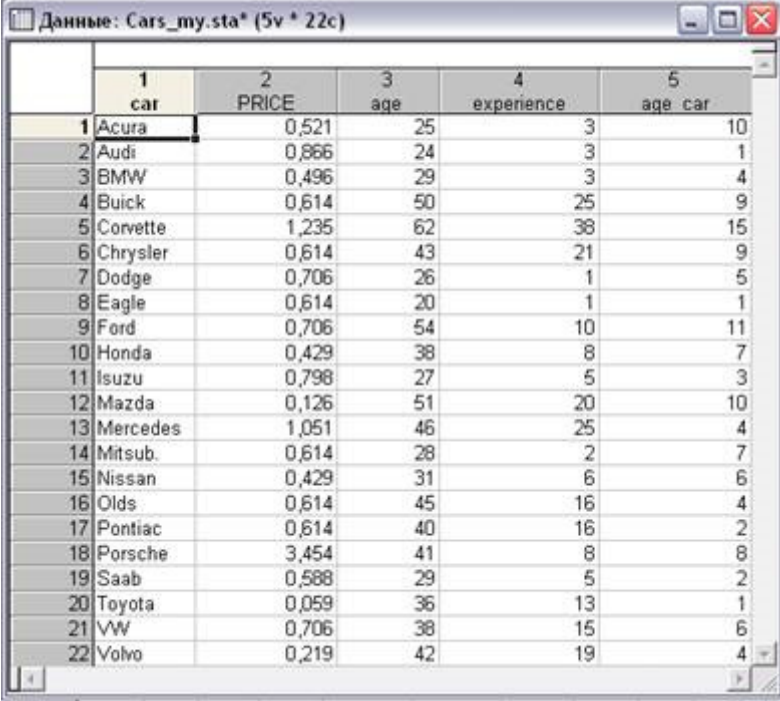
Данные могут поступать как в исходном виде, так и в виде матрицы расстояний между объектами.

Наблюдения и переменные можно кластеризовать, используя различные меры расстояния (евклидово, квадрат евклидова, манхэттенское, Чебышева и др.) и различные правила объединения кластеров (одиночная, полная связь, невзвешенное и взвешенное попарное среднее по группам и др.).

Постановка задачи

Исходный файл данных содержит следующую информацию об автомобилях и их владельцах (cars_my.sta):

- марка автомобиля – первая переменная;
- стоимость автомобиля – вторая переменная;
- возраст водителя – третья переменная;
- стаж водителя – четвертая переменная;
- возраст автомобиля – пятая переменная;



Данные: Cars_my.sta* (5v * 22c)

| | 1 | 2 | 3 | 4 | 5 |
|----|----------|-------|-----|------------|---------|
| | car | PRICE | age | experience | age car |
| 1 | Acura | 0,521 | 25 | 3 | 10 |
| 2 | Audi | 0,866 | 24 | 3 | 1 |
| 3 | BMW | 0,496 | 29 | 3 | 4 |
| 4 | Buick | 0,614 | 50 | 25 | 9 |
| 5 | Corvette | 1,235 | 62 | 38 | 15 |
| 6 | Chrysler | 0,614 | 43 | 21 | 9 |
| 7 | Dodge | 0,706 | 26 | 1 | 5 |
| 8 | Eagle | 0,614 | 20 | 1 | 1 |
| 9 | Ford | 0,706 | 54 | 10 | 11 |
| 10 | Honda | 0,429 | 38 | 8 | 7 |
| 11 | Isuzu | 0,798 | 27 | 5 | 3 |
| 12 | Mazda | 0,126 | 51 | 20 | 10 |
| 13 | Mercedes | 1,051 | 46 | 25 | 4 |
| 14 | Mitsub. | 0,614 | 28 | 2 | 7 |
| 15 | Nissan | 0,429 | 31 | 6 | 6 |
| 16 | Olds | 0,614 | 45 | 16 | 4 |
| 17 | Pontiac | 0,614 | 40 | 16 | 2 |
| 18 | Porsche | 3,454 | 41 | 8 | 8 |
| 19 | Saab | 0,588 | 29 | 5 | 2 |
| 20 | Toyota | 0,059 | 36 | 13 | 1 |
| 21 | VW | 0,706 | 38 | 15 | 6 |
| 22 | Volvo | 0,219 | 42 | 19 | 4 |

Целью данного анализа является разбиение автомобилей и их владельцев на классы, каждый из которых соответствует определенной рискованной группе. Наблюдения, попавшие в одну группу, характеризуются одинаковой вероятностью наступления страхового случая, которая впоследствии оценивается страховщиком.

Использование кластер-анализа для решения данной задачи наиболее эффективно. В общем случае кластер-анализ предназначен для объединения некоторых объектов в классы (кластеры) таким образом, чтобы в один класс попадали максимально схожие, а объекты различных классов максимально отличались друг от друга. Количественный показатель сходства рассчитывается заданным способом на основании данных, характеризующих объекты.

Масштаб измерений

Все кластерные алгоритмы нуждаются в оценках расстояний между кластерами или объектами, и ясно, что при вычислении расстояния необходимо задать масштаб измерений.

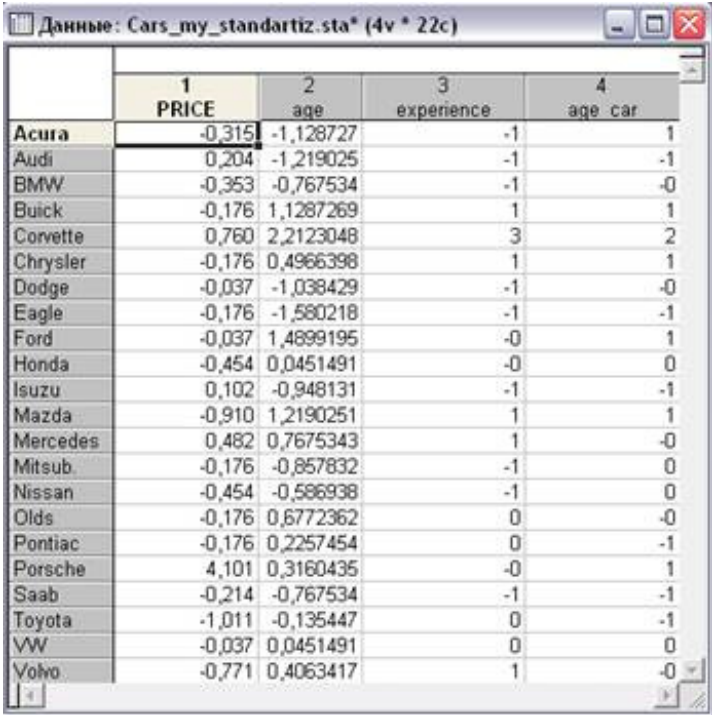
Поскольку различные измерения используют абсолютно различные типы шкал, данные необходимо стандартизовать. Предварительно выполним вспомогательные преобразования файла исходных данных: в меню *Данные* выберите пункт *Менеджер имен строк (Case Name Manager)* затем в опции *Variable* ввести имя переменной *car*, нажать *Ок*. В результате этих действий строки таблицы приобретут названия, как в поле *car*.

Удаляем переменную (колонку) *car*.

Для стандартизации в меню *Данные* выберите пункт *Стандартизовать*, после чего каждая переменная будет иметь среднее 0 и стандартное отклонение 1.

Сохраняем данные под именем *cars_my_standartiz.sta* и далее работаем с ним.

Таблица со стандартизованными переменными приведена ниже.

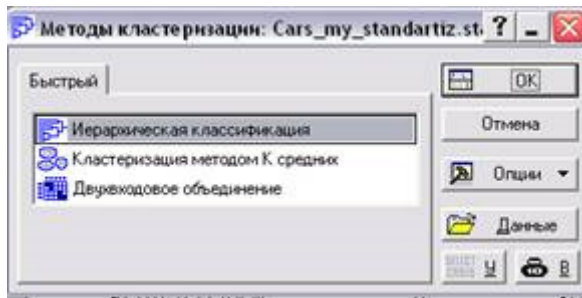


| | 1 PRICE | 2 age | 3 experience | 4 age car |
|----------|------------|-----------|-----------------|--------------|
| Acura | -0,315 | -1,128727 | -1 | 1 |
| Audi | 0,204 | -1,219025 | -1 | -1 |
| BMW | -0,353 | -0,767534 | -1 | -0 |
| Buick | -0,176 | 1,1287269 | 1 | 1 |
| Corvette | 0,760 | 2,2123048 | 3 | 2 |
| Chrysler | -0,176 | 0,4966398 | 1 | 1 |
| Dodge | -0,037 | -1,038429 | -1 | -0 |
| Eagle | -0,176 | -1,580218 | -1 | -1 |
| Ford | -0,037 | 1,4899195 | -0 | 1 |
| Honda | -0,454 | 0,0451491 | -0 | 0 |
| Isuzu | 0,102 | -0,948131 | -1 | -1 |
| Mazda | -0,910 | 1,2190251 | 1 | 1 |
| Mercedes | 0,482 | 0,7675343 | 1 | -0 |
| Mitsub | -0,176 | -0,857832 | -1 | 0 |
| Nissan | -0,454 | -0,586938 | -1 | 0 |
| Olds | -0,176 | 0,6772362 | 0 | -0 |
| Pontiac | -0,176 | 0,2257454 | 0 | -1 |
| Porsche | 4,101 | 0,3160435 | -0 | 1 |
| Saab | -0,214 | -0,767534 | -1 | -1 |
| Toyota | -1,011 | -0,135447 | 0 | -1 |
| VW | -0,037 | 0,0451491 | 0 | 0 |
| Volvo | -0,771 | 0,4063417 | 1 | -0 |

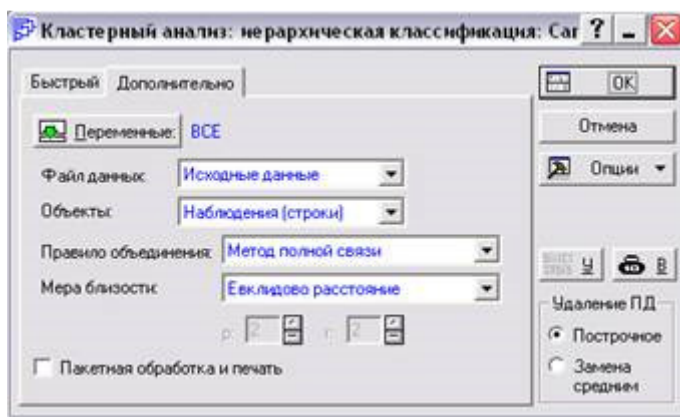
Шаг 1. Иерархическая классификация

На первом этапе выясним, формируют ли автомобили "естественные" кластеры, которые могут быть осмыслены.

Выберем *Кластерный анализ* в меню *Анализ (Statistics)* - *Многомерный разведочный анализ* для отображения стартовой панели модуля *Кластерный анализ*. В этом диалоге выберем *Иерархическая классификация* и нажмем *ОК*.



Нажмем кнопку *Переменные*, выберем *Все*, в поле *Объекты* выберем *Наблюдения (строки)*. В качестве правила объединения отметим *Метод полной связи*, в качестве меры близости – *Евклидово расстояние*. Нажмем *ОК*.

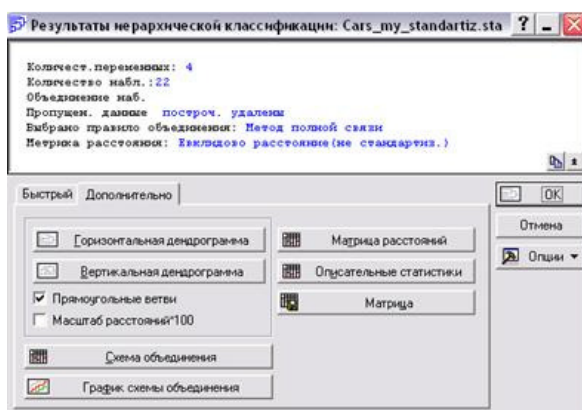


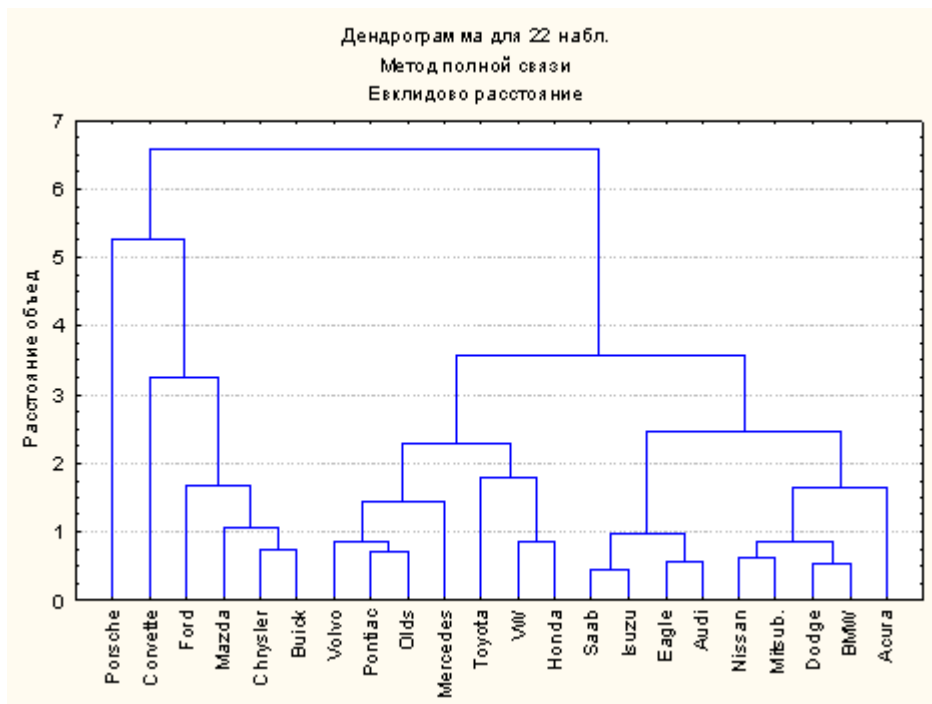
Метод полной связи определяет расстояние между кластерами как наибольшее расстояние между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями").

Мера близости, определяемая евклидовым расстоянием, является геометрическим расстоянием в n- мерном пространстве и вычисляется следующим образом:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Наиболее важным результатом, получаемым в результате древовидной кластеризации, является иерархическое дерево. Нажмем на кнопку *Вертикальная дендрограмма*.





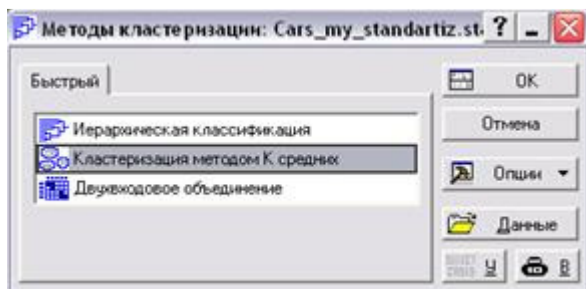
Вначале древовидные диаграммы могут показаться немного запутанными, однако после некоторого изучения они становятся более понятными. Диаграмма начинается сверху (для вертикальной дендрограммы) с каждого автомобиля в своем собственном кластере.

Как только вы начнете двигаться вниз, автомобили, которые "теснее соприкасаются друг с другом" объединяются и формируют кластеры. Каждый узел диаграммы, приведенной выше, представляет объединение двух или более кластеров, положение узлов на вертикальной оси определяет расстояние, на котором были объединены соответствующие кластеры.

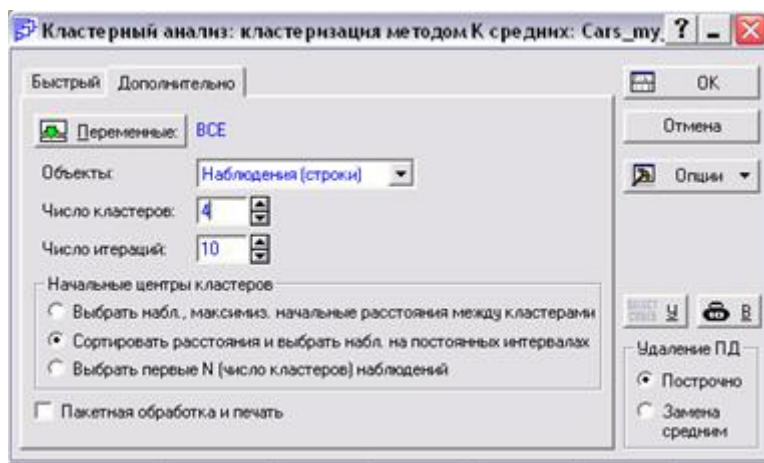
Шаг 2. Кластеризация методом К средних

Исходя из визуального представления результатов, можно сделать предположение, что автомобили образуют четыре естественных кластера. Проверим данное предположение, разбив исходные данные методом К средних на 4 кластера, и проверим значимость различия между полученными группами.

В Стартовой панели модуля *Кластерный анализ* выберем *Кластеризация методом К средних*.



Нажмем кнопку *Переменные* и выберем *Все*, в поле *Объекты* выберем *Наблюдения (строки)*, зададим 4 кластера разбиения.



Метод *K-средних* заключается в следующем: вычисления начинаются с k случайно выбранных наблюдений (в нашем случае $k=4$), которые становятся центрами групп, после чего объектный состав кластеров меняется с целью минимизации изменчивости внутри кластеров и максимизации изменчивости между кластерами.

Каждое следующее наблюдение ($K+1$) относится к той группе, мера сходства с центром тяжести которого минимальна.

После изменения состава кластера вычисляется новый центр тяжести, чаще всего как вектор средних по каждому параметру. Алгоритм продолжается до тех пор, пока состав кластеров не перестанет меняться.

Когда результаты классификации получены, можно рассчитать среднее значение показателей по каждому кластеру, чтобы оценить, насколько они различаются между собой.

В окне *Результаты метода K средних* выберем *Дисперсионный анализ (Analysis of Variance)* для определения значимости различия между полученными кластерами.

| перемен. | Дисперсионный анализ (Cars_my_standartiz.sta) | | | | | |
|------------|-----------------------------------------------|----|-----------|----|----------|-----------|
| | Между SS | сс | Внутри SS | сс | F | значим. р |
| PRICE | 17,75805 | 3 | 3,241945 | 18 | 32,86555 | 0,000000 |
| age | 18,05119 | 3 | 2,948812 | 18 | 36,72906 | 0,000000 |
| experience | 14,71184 | 3 | 6,288156 | 18 | 14,03767 | 0,000058 |
| age_car | 12,24617 | 3 | 8,753834 | 18 | 8,39369 | 0,001058 |

Итак, значение $p < 0.05$, что говорит о значимом различии.

Нажмем кнопку *Элементы кластеров и расстояния* для просмотра наблюдений, входящих в каждый из кластеров. Опция также позволяет отобразить евклидовы расстояния объектов от центров (средних значений) соответствующих им кластеров.

Первый кластер:

| | Элементы кластера номер 1 (Cars_my_standartiz.sta) и расстояния до центра кластера. Кластер содержит 1 набл. | |
|--------|--------------------------------------------------------------------------------------------------------------|------|
| | Porsche | |
| Расст. | | 0,00 |

Второй кластер:

| | | | | | | | |
|--------------------------------------------------------------------------------------------------------------------|----------|----------|----------|----------|----------|----------|----------|
| Элементы кластера номер 2 (Cars_my_standartiz.sta) и расстояния до центра кластера. Кластер содержит 7 набл. | | | | | | | |
| | Honda | Mercedes | Olds | Pontiac | Toyota | VW | Volvo |
| Расст. | 0,590329 | 0,661432 | 0,204147 | 0,276696 | 0,594585 | 0,327201 | 0,284429 |

Третий кластер:

| | | | | | | | | | |
|--------------------------------------------------------------------------------------------------------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| Элементы кластера номер 3 (Cars_my_standartiz.sta) и расстояния до центра кластера. Кластер содержит 9 набл. | | | | | | | | | |
| | Acura | Audi | BMW | Dodge | Eagle | Isuzu | Mitsub. | Nissan | Saab |
| Расст. | 0,763888 | 0,494066 | 0,154399 | 0,158544 | 0,546472 | 0,239325 | 0,367393 | 0,363288 | 0,343910 |

Четвертый кластер:

| | | | | | |
|--------------------------------------------------------------------------------------------------------------------|----------|----------|----------|----------|----------|
| Элементы кластера номер 4 (Cars_my_standartiz.sta) и расстояния до центра кластера. Кластер содержит 5 набл. | | | | | |
| | Buick | Corvette | Chrysler | Ford | Mazda |
| Расст. | 0,282375 | 1,145248 | 0,482189 | 0,662676 | 0,441467 |

Итак, в каждом из четырех кластеров находятся объекты со схожим влиянием на процесс убытков.

Шаг 3. Описательные статистики

Знание описательных статистик в каждой группе, безусловно, является важным для любого исследователя.

Отображение статистик для каждого кластера из диалога Результаты метода К средних в данном случае не представляет интереса, т.к. данные были стандартизованы.

Нажмем кнопку *Сохранить классификацию и расстояния*.

Данные: Таблица данных14* (7v * 22с)

| Cars_my_standartiz.sta | | | | | | | |
|------------------------|--------|----------|------------|---------|---------|---------|--------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| | PRICE | age | experience | age_car | НАБЛ.НО | КЛАСТЕР | РАССТ. |
| Acura | -0,315 | -1,12873 | -1 | 1 | 1 | 3 | 0,76 |
| Audi | 0,204 | -1,21903 | -1 | -1 | 2 | 3 | 0,49 |
| BMW | -0,353 | -0,76753 | -1 | -0 | 3 | 3 | 0,15 |
| Buick | -0,176 | 1,128727 | 1 | 1 | 4 | 4 | 0,28 |
| Corvette | 0,760 | 2,212305 | 3 | 2 | 5 | 4 | 1,15 |
| Chrysler | -0,176 | 0,49664 | 1 | 1 | 6 | 4 | 0,48 |
| Dodge | -0,037 | -1,03843 | -1 | -0 | 7 | 3 | 0,16 |
| Eagle | -0,176 | -1,58022 | -1 | -1 | 8 | 3 | 0,55 |
| Ford | -0,037 | 1,48992 | -0 | 1 | 9 | 4 | 0,66 |
| Honda | -0,454 | 0,045149 | -0 | 0 | 10 | 2 | 0,59 |
| Isuzu | 0,102 | -0,94813 | -1 | -1 | 11 | 3 | 0,24 |
| Mazda | -0,910 | 1,219025 | 1 | 1 | 12 | 4 | 0,44 |
| Mercedes | 0,482 | 0,767534 | 1 | -0 | 13 | 2 | 0,65 |
| Mitsub. | -0,176 | -0,85783 | -1 | 0 | 14 | 3 | 0,37 |
| Nissan | -0,454 | -0,58694 | -1 | 0 | 15 | 3 | 0,36 |
| Olds | -0,176 | 0,677236 | 0 | -0 | 16 | 2 | 0,20 |
| Pontiac | -0,176 | 0,225745 | 0 | -1 | 17 | 2 | 0,28 |
| Porsche | 4,101 | 0,316044 | -0 | 1 | 18 | 1 | 0,00 |
| Saab | -0,214 | -0,76753 | -1 | -1 | 19 | 3 | 0,34 |
| Toyota | -1,011 | -0,13545 | 0 | -1 | 20 | 2 | 0,59 |
| VW | -0,037 | 0,045149 | 0 | 0 | 21 | 2 | 0,33 |
| Volvo | -0,771 | 0,406342 | 1 | -0 | 22 | 2 | 0,28 |

Таблица стандартизованных данных дополнилась информацией о кластере, к которому принадлежит наблюдение, евклидовом расстоянии и номере наблюдения.

Скопируем переменную КЛАСТЕР в исходную таблицу данных.

Данные: Cars_my_sta* (6v * 22c)

| | 1 | 2 | 3 | 4 | 5 | 6 |
|----|----------|-------|-----|------------|---------|---------|
| | car | PRICE | age | experience | age car | КЛАСТЕР |
| 1 | Acura | 0,521 | 25 | 3 | 10 | 3 |
| 2 | Audi | 0,866 | 24 | 3 | 1 | 3 |
| 3 | BMW | 0,496 | 29 | 3 | 4 | 3 |
| 4 | Buick | 0,614 | 50 | 25 | 9 | 4 |
| 5 | Corvette | 1,235 | 62 | 38 | 15 | 4 |
| 6 | Chrysler | 0,614 | 43 | 21 | 9 | 4 |
| 7 | Dodge | 0,706 | 26 | 1 | 5 | 3 |
| 8 | Eagle | 0,614 | 20 | 1 | 1 | 3 |
| 9 | Ford | 0,706 | 54 | 10 | 11 | 4 |
| 10 | Honda | 0,429 | 38 | 8 | 7 | 2 |
| 11 | Isuzu | 0,798 | 27 | 5 | 3 | 3 |
| 12 | Mazda | 0,126 | 51 | 20 | 10 | 4 |
| 13 | Mercedes | 1,051 | 46 | 25 | 4 | 2 |
| 14 | Mitsub. | 0,614 | 28 | 2 | 7 | 3 |
| 15 | Nissan | 0,429 | 31 | 6 | 6 | 3 |
| 16 | Olds | 0,614 | 45 | 16 | 4 | 2 |
| 17 | Pontiac | 0,614 | 40 | 16 | 2 | 2 |
| 18 | Porsche | 3,454 | 41 | 8 | 8 | 1 |
| 19 | Saab | 0,588 | 29 | 5 | 2 | 3 |
| 20 | Toyota | 0,059 | 36 | 13 | 1 | 2 |
| 21 | VW | 0,706 | 38 | 15 | 6 | 2 |
| 22 | Volvo | 0,219 | 42 | 19 | 4 | 2 |

Теперь для каждого кластера можно вычислить основные описательные статистики.

В меню *Анализ – Основные статистики и таблицы* выберем опцию *Группировка и Однофакторный Дисперсионный анализ (Breakdown & one-way ANOVA)*.

Statistics by Groups (Breakdown): Cars_my_with_sluster.sta

Individual tables | Lists of tables

Variables

Dependent: none

Grouping: none

Codes for grouping variables: none

OK

Cancel

Options

SELECT CASES

Weighted moments

DF =

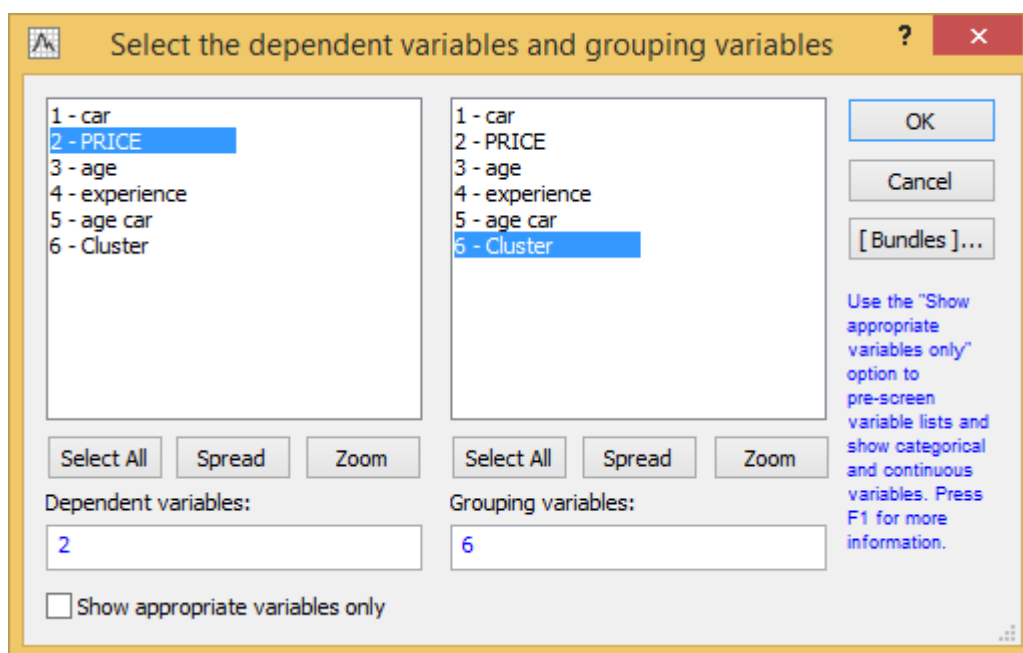
☒ W-1 ☐ N-1

MD deletion

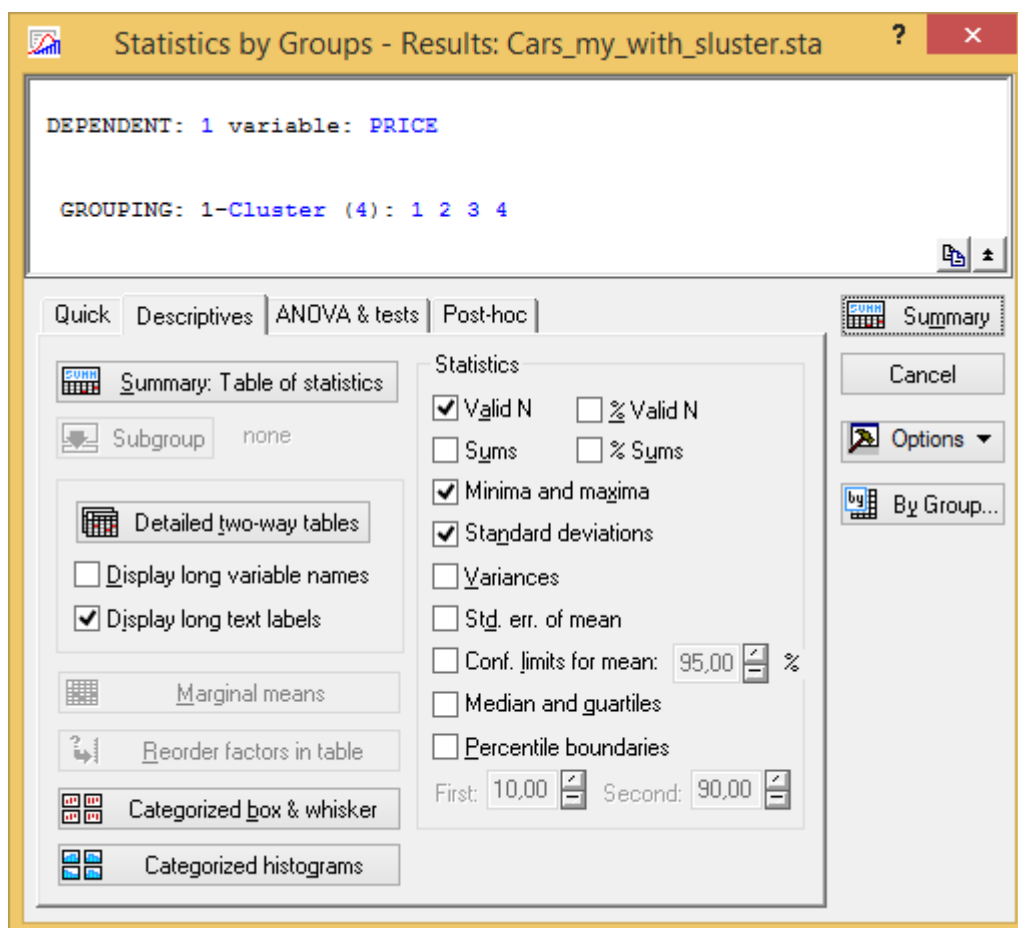
☐ Casewise

☒ Pairwise

В опции *Variables* определяем исследуемые переменные:



Далее (после Ок):



Далее жмем Summary: Table of statistics, получаем результат исследований по переменной PRICE (см. таблицу ниже). Повторяя процесс для каждой переменной, получаем таблицы описательных статистик для каждого из показателей.

Ниже приведены таблицы описательных статистик для каждого из показателей:

| Итоговая таблица средних (Cars_my.sta) N=22 (Нет пропусков в завис. перем.) | | | | | |
|--------------------------------------------------------------------------------|------------------|------------|-------------------|------------------|-------------------|
| КЛАСТЕР | PRICE Среднее | PRICE N | PRICE Ст.откл. | PRICE Минимум | PRICE Максимум |
| 1 | 3,454206 | 1 | 0,000000 | 3,454206 | 3,454206 |
| 2 | 0,527076 | 7 | 0,327905 | 0,058831 | 1,050549 |
| 3 | 0,625660 | 9 | 0,142386 | 0,428624 | 0,865652 |
| 4 | 0,658904 | 5 | 0,394541 | 0,126066 | 1,235446 |
| Всего | 0,730418 | 22 | 0,664142 | 0,058831 | 3,454206 |

| Итоговая таблица средних (Cars_my.sta) N=22 (Нет пропусков в завис. перем.) | | | | | |
|--------------------------------------------------------------------------------|----------------|----------|-----------------|----------------|-----------------|
| КЛАСТЕР | age Среднее | age N | age Ст.откл. | age Минимум | age Максимум |
| 1 | 41,00000 | 1 | 0,00000 | 41,00000 | 41,00000 |
| 2 | 40,71429 | 7 | 3,77334 | 36,00000 | 46,00000 |
| 3 | 26,55556 | 9 | 3,28295 | 20,00000 | 31,00000 |
| 4 | 52,00000 | 5 | 6,89202 | 43,00000 | 62,00000 |
| Всего | 37,50000 | 22 | 11,07442 | 20,00000 | 62,00000 |

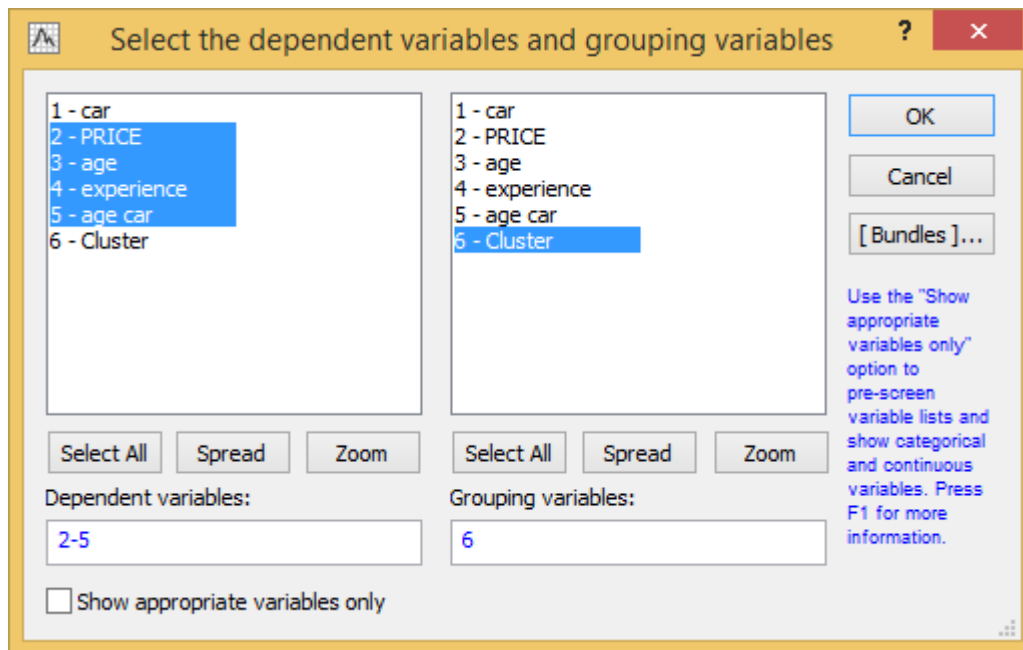
| Итоговая таблица средних (Cars_my.sta) N=22 (Нет пропусков в завис. перем.) | | | | | |
|--------------------------------------------------------------------------------|-----------------------|-----------------|------------------------|-----------------------|------------------------|
| КЛАСТЕР | experience Среднее | experience N | experience Ст.откл. | experience Минимум | experience Максимум |
| 1 | 8,00000 | 1 | 0,00000 | 8,00000 | 8,00000 |
| 2 | 16,00000 | 7 | 5,22813 | 8,00000 | 25,00000 |
| 3 | 3,22222 | 9 | 1,78730 | 1,00000 | 6,00000 |
| 4 | 22,80000 | 5 | 10,13410 | 10,00000 | 38,00000 |
| Всего | 11,95455 | 22 | 9,77108 | 1,00000 | 38,00000 |

| Итоговая таблица средних (Cars_my.sta) N=22 (Нет пропусков в завис. перем.) | | | | | |
|--------------------------------------------------------------------------------|--------------------|--------------|---------------------|--------------------|---------------------|
| КЛАСТЕР | age_car Среднее | age_car N | age_car Ст.откл. | age_car Минимум | age_car Максимум |
| 1 | 8,00000 | 1 | 0,000000 | 8,000000 | 8,00000 |
| 2 | 4,00000 | 7 | 2,081666 | 1,000000 | 7,00000 |
| 3 | 4,33333 | 9 | 3,000000 | 1,000000 | 10,00000 |
| 4 | 10,80000 | 5 | 2,489980 | 9,000000 | 15,00000 |
| Всего | 5,86364 | 22 | 3,745416 | 1,000000 | 15,00000 |

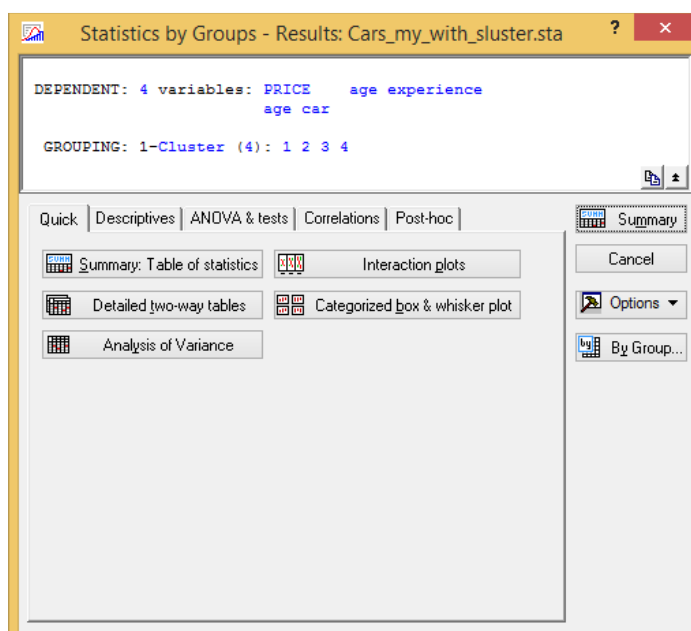
Знание основных описательных статистик в каждом кластере может быть использовано экспертом для оценки убытков страховой компании.

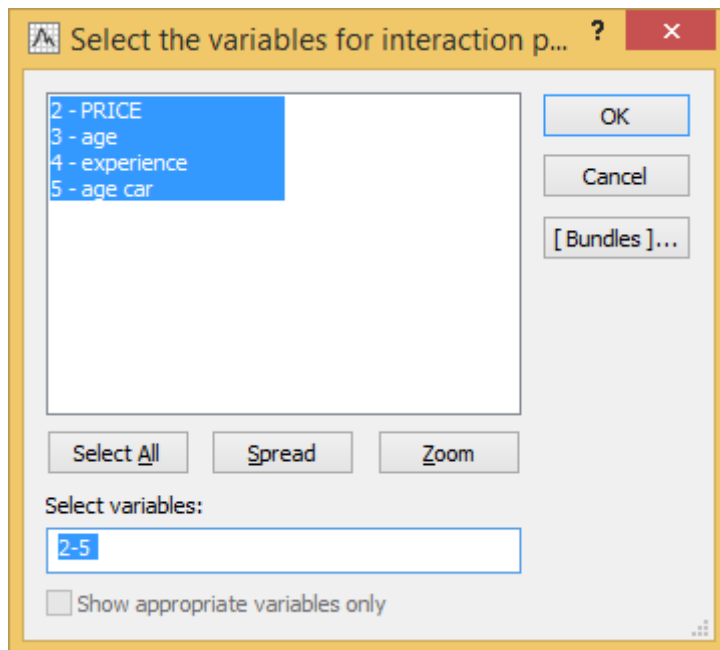
Построим график средних и доверительных интервалов для переменных в каждом кластере. Для этого в меню *Анализ – Основные статистики и таблицы* заново выберем опцию *Группировка и Однофакторный Дисперсионный анализ (Breakdown & one-way ANOVA)*.

В опции *Variables* определяем исследуемые переменные:

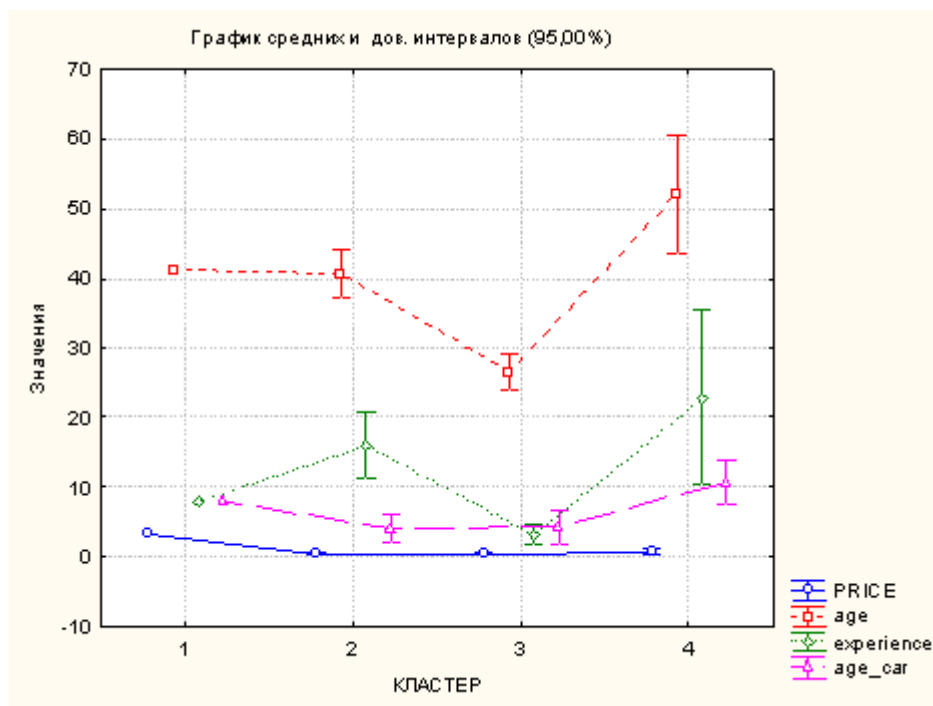


Жмем ОК. Далее рисуем графики нажав на Interaction Plots:





Итог:



Итак, имея полученные результаты, для каждого кластера экспертом может быть определена вероятность наступления страхового случая.

6 Контрольные вопросы

1. Дать понятие кластерного анализа (cluster analysis)
2. Дать определение понятию кластер
3. Каковы задачи кластерного анализа?
4. Что является результатом решением задачи кластерного анализа?

5. Что есть целевая функция кластеризации?
6. Что представляет собой целевая функция в виде внутригрупповой суммы квадратов отклонения (привести выражение, дать пояснения)?
7. Дать определение центра кластера, радиуса кластера, размера кластера, спорного объекта
8. Дать определение и обоснование необходимости стандартизация (standardization) или нормирование (normalization) переменных
9. Дать выражение для вычисления евклидова расстояния как метрики в кластерном анализе
10. Что из себя представляют метрики: квадрат евклидова расстояния, обобщенное степенное расстояние, расстояние Чебышева, Манхэттенское расстояние, процент несогласия?
11. Иерархические агломеративные методы (Agglomerative Nesting, AGNES) – дать характеристику
12. Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA) – дать характеристику
13. Дендрограмма – дать определение, характеристику и правила построения
14. Что из себя представляет обобщенная агломеративная процедура построения дендрограммы?
15. Расстояния между кластерами: расстояние “Ближайшего соседа”, расстояние “Дальнего соседа” – дать определение
16. Расстояния между кластерами: невзвешенное попарное среднее, взвешенное попарное среднее – дать определение и особенности
17. Расстояния между кластерами: невзвешенный центроидный метод, взвешенный центроидный метод (медиана) – дать определение и особенности
18. Расстояния между кластерами: метод Варда (Ward) – дать определение и особенности
19. Дать математическое описание алгоритма метода k - средних