КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ – это метод классификационного анализа; его основное назначение – разбиение множества исследуемых объектов и признаков на однородные в некотором смысле группы, или кластеры. Это многомерный статистический метод, поэтому предполагается, что исходные данные могут быть значительного объема, т.е. существенно большим может быть как количество объектов исследования (наблюдений), так и признаков, характеризующих эти объекты.

В прикладной статистике многомерными статистическими методами долго не могли пользоваться из-за отсутствия вычислительной техники для обработки больших массивов данных. Активно эти методы стали развиваться со второй половины XX в. при появлении быстродействующих компьютеров, выполняющих за доли секунды необходимые вычисления, на которые до этого уходили дни, недели, месяцы [13]. В настоящее время препятствием к широкому использованию многомерных статистических методов, в том числе и кластерного анализа, является отсутствие у исследователей навыков и умения работать со статистическими пакетами прикладных программ.

Техника кластеризации может применяться в самых различных прикладных областях, в том числе и в медицине. Например, кластеризация заболеваний, симптомов, признаков заболеваний, методов лечения может привести к более полному и глубокому пониманию медицинских проблем, связанных с лечением больных.

Большое достоинство кластерного анализа в том, что он дает возможность производить разбиение объектов не по одному признаку, а по ряду признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет исследовать множество исходных данных практически произвольной природы.

Так как кластеры – это группы однородности, то задача кластерного анализа заключается в том, чтобы на основании признаков объектов разбить их множество на m (m – целое) кластеров так, чтобы каждый объект принадлежал только одной группе разбиения. При этом объекты, принадлежащие одному кластеру, должны быть однородными (сходными), а объекты, принадлежащие разным кластерам, – разнородными.

Если объекты кластеризации представить как точки в *n-мерном* пространстве признаков (n – количество признаков, характеризующих объекты), то сходство между объектами определяется через понятие расстояния между точками, так как интуитивно понятно, что чем меньше расстояние между объектами, тем они более схожи.

Перечислим используемые в модуле *Кластерный анализ* программы STATISTICA функции расстояний (метрики) [21].

Евклидово расстояние – наиболее популярная метрика, является геометрическим расстоянием в многомерном пространстве. Данная метрика, как и большинство других, чувствительна к изменению единиц измерения осей. Например, если сантиметры перевести в миллиметры, то изменится и исчисляемое расстояние. Поэтому при использовании большинства метрик кластерный анализ предполагает предварительную стандартизацию (нормирование, см. 1.9) исходных данных.

Квадрат евклидова расстояния используют, если необходимо придать большие веса более отдаленным друг от друга объектам.

Манхэттенское расстояние городских кварталов уменьшает влияние отдельных больших разностей между одноименными координатами точек, так как при вычислении расстояния эти разности не возводятся в квадрат (в отличие от евклидовой метрики).

Расстояние Чебышева применяют, когда желают определить два объекта как различные, если они различаются по какой-либо одной координате.

Расстояние Минковского используют, когда надо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются.

Процент несогласия применяют в тех случаях, когда данные являются категориальными.

Все приведенные расстояния пригодны, если объекты кластеризации можно представить как точки в *n-мерном* пространстве. При решении большого количества задач объекты нельзя представить как точки в *n-мерном* пространстве. В этом случае целесообразно в качестве расстояния использовать метрику *1 – коэффициент корреляции Пирсона***.**

Алгоритмов кластерного анализа достаточно много. Все их можно подразделить на иерархические и неиерархические.

Иерархические (древовидные) процедуры – наиболее распространённые алгоритмы кластерного анализа по их реализации на ЭВМ. Различают агломеративные (от слова *agglomerate* – собирать) и итеративные дивизивные (от слова *division* – разделять) процедуры.

Принцип работы иерархических агломеративных процедур состоит в последовательном объединении групп элементов сначала самых близких, а затем всё более отдалённых друг от друга. Принцип работы иерархических дивизивных процедур, наоборот, состоит в последовательном разделении групп элементов сначала самых далёких, а затем всё более близких друг от друга. Большинство этих алгоритмов исходит из матрицы расстояний (сходства). К недостаткам иерархических процедур следует отнести громоздкость их вычислительной реализации. На каждом шаге алгоритмы требуют вычисления матрицы расстояний, а следовательно, ёмкой машинной памяти и большого количества времени. В этой связи реализация таких алгоритмов при числе наблюдений, большем нескольких сотен, нецелесообразна, а в ряде случаев и невозможна.

Общий принцип работы агломеративного алгоритма следующий. На первом шаге каждое наблюдение рассматривается как отдельный кластер. В дальнейшем на каждом шаге работы алгоритма происходит объединение двух самых близких кластеров, и с учётом принятого расстояния по формуле пересчитывается матрица расстояний, размерность которой, очевидно, снижается на единицу. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс. Большинство программ, реализующих алгоритм иерархической классификации, предусматривает графическое представление классификации в виде дендрограммы.

В программе STATISTICA реализованы агломеративные методы минимальной дисперсии – древовидная кластеризация и двухвходовая кластеризация, а также дивизивный метод *k-средних*.

В методе древовидной кластеризации предусмотрены различные правила иерархического объединения в кластеры [22]:

1. Правило *одиночной связи*. На первом шаге объединяются два наиболее близких объекта, т.е. имеющие максимальную меру сходства. На следующем шаге к ним присоединяется объект с максимальной мерой сходства с одним из объектов кластера, т.е. для его включения в кластер требуется максимальное сходство лишь с одним членом кластера. Метод называют еще методом ближайшего соседа, так как расстояние между двумя кластерами определяется как расстояние между двумя наиболее близкими объектами в различных кластерах. Это правило «нанизывает» объекты для формирования кластеров. Недостаток данного метода – образование слишком больших продолговатых кластеров.

2. Правило *полных связей*. Метод позволяет устранить недостаток, присущий методу одиночной связи. Суть правила в том, что два объекта, принадлежащих одной и той же группе (кластеру), имеют коэффициент сходства, который больше некоторого порогового значения *S*. В терминах евклидова расстояния это означает, что расстояние между двумя точками (объектами) кластера не должно превышать некоторого порогового значения *d*. Таким образом, *d* определяет максимально допустимый диаметр подмножества, образующего кластер. Этот метод называют еще методом наиболее удаленных соседей, так

как при достаточно большом пороговом значении d расстояние между кластерами определяется наибольшим расстоянием между любыми двумя объектами в различных кластерах.

3. Правило *невзвешенного попарного среднего*. Расстояние между двумя кластерами определяется как среднее расстояние между всеми парами объектов в них. Метод эффективен, когда объекты в действительности формируют различные группы, однако он работает одинаково хорошо и в случаях протяженных (цепочного типа) кластеров.

4. Правило *взвешенное попарное среднее*. Метод идентичен предыдущему, за исключением того, что при вычислении размер соответствующих кластеров используется в качестве весового коэффициента. Желательно этот метод использовать, когда предполагаются неравные размеры кластеров.

5. *Невзвешенный центроидный* метод. Расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

6. *Взвешенный центроидный* метод. Идентичен предыдущему, за исключением того, что при вычислениях расстояния используют веса для учета разности между размерами кластеров. Поэтому, если имеются (или подозреваются) значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее предыдущего.

7. Правило *Уорда* (Варда). В этом методе в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений, которая есть не 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 отклонений. Этот метод направлен на объединение близко расположенных кластеров. Замечено, что метод Уорда приводит к образованию кластеров примерно равных размеров и имеющих форму гиперсфер.

Ранее мы рассмотрели методы кластеризации объектов (наблюдений), однако иногда кластеризация по переменным может привести к достаточно интересным результатам. В модуле *Кластерный анализ* также предусмотрена эффективная двухвходовая процедура, которая позволит кластеризовать сразу в двух направлениях – по наблюдениям и переменным.

Метод k-средних

Предположим, есть гипотезы относительно числа m кластеров (по переменным или наблюдениям). Тогда можно задать программе создать ровно m кластеров так, чтобы они были настолько различны, насколько это возможно. Именно для решения задач этого типа предназначен метод *k-means* (*k-средних*). Гипотеза может основываться на теоретических соображениях, результатах предшествующих исследований или догадке. Выполняя последовательное разбиение на различное число кластеров, можно сравнивать качество получаемых решений.

Программа начинает с m случайно выбранных кластеров, а затем изменяет принадлежность объектов к ним, чтобы минимизировать изменчивость внутри кластеров и максимизировать изменчивость между кластерами. Алгоритм случайным образом в пространстве назначает центры будущих кластеров. Затем вычисляет расстояние между центрами кластеров и каждым объектом, и объект приписывается к тому кластеру, к которому он ближе всего. Завершив приписывание, алгоритм вычисляет средние значения для каждого кластера. Этих средних будет столько, сколько используется переменных для проведения анализа, – k штук. Набор средних представляет собой координаты нового положения центра кластера. Алгоритм вновь вычисляет расстояние от каждого объекта до центров кластеров и приписывает объекты к ближайшему кластеру. Вновь вычисляются центры тяжести кластеров, и этот процесс повторяется до тех пор, пока центры тяжести не перестанут «мигрировать» в пространстве.

Если в древовидной кластеризации можно использовать категориальные переменные, то так как в методе *k-средних* в качестве метрики используют евклидову метрику, то перед проведением кластеризации необходимо стандартизовать переменные. По этой же причине в методе предполагается, что переменные непрерывные и измерены как минимум в интервальной шкале.

Разбиение модельной базы предприятий на классы методом кластерного анализа

Наша задача кластерного анализа сводится к разбиению множемтва элементов корреляционной матрицы признаков (данные по 1000 предприятям) на *m* групп (5 кластеров) удовлетворяющих критерию оптимальности:

- каждое предприятие должно принадлежать одному и только одному подмножеству разбиения (кластеру);
- предприятия, принадлежащие одному и тому же кластеру, должны быть сходными;
- предприятия, принадлежащие разным кластерам, должны быть разнородными.

Этот критерий может представлять собой ***целевую функцию*** – функционал, выражающий уровни желательности различных разбиений и группировок. Сходство и разнородность объектов будем определять по средствам *m*-мерного евклидового расстояния между векторами измерений.

Запустим модуль ***Cluster Analysis***(кластерный анализ) через меню Statistics – Multivariate Exploratory Techniques (многомерные исследовательские методы). Откроется стартовая панель модуля. На вкладке Quick находится список методов кластерного анализа, реализованных в программе STATISTICA 6. (рис.1)

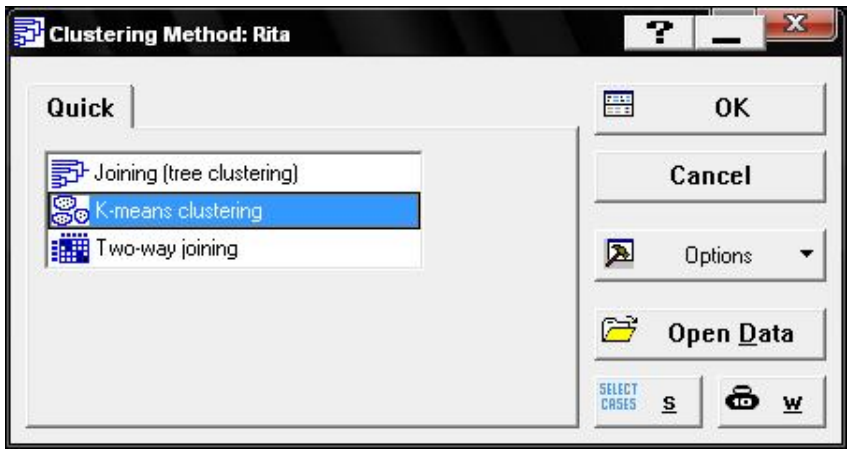

Рисунок 1. Диалоговое окно модуля ***Cluster Analysis***

Это Joining tree clustering (древовидная кластеризация), k-means clustering (метод k-средних) и Two-way joining (двухвходовая кластеризация). Выберем ***k-means clustering*** (метод k-средних).

Предполагаем, что есть гипотезы относительно числа *m=5* групп/кластеров по 100 наблюдениям/предприятиям. Тогда задаем программе создание ровно *m=5* кластеров так, чтобы они были настолько различны насколько это возможно. Гипотеза может основываться на теоретических соображениях, результатах предшествующих исследований или догадке. Или можно определить число кластеров опытным путем: выполняя последовательное разбиение на различное число кластеров *m=1…5* можно сравнивать качество получаемых решений. Так предполагается, что показатели предприятий попадают в основном в пять различных категорий.

При помощи модуля убедимся, действительно ли это так. Алгоритм, реализованный в программе Statistica 6. сводится к следующим шагам:

Шаг 1. Алгоритм случайным образом в пространстве назначает центры будущих кластеров.

Шаг 2. Программа вычисляет расстояние между центрами кластеров и каждым объектом, и объект приписывается к тому кластеру, к которому он ближе всего.

Шаг 3. После приписывания, алгоритм вычисляет средние значения для каждого кластера. Этих средних будет столько, сколько используется переменных для проведения анализа, *k* штук. Набор средних представляет собой координаты нового положения центра кластера.

Шаг 4. Алгоритм вновь вычисляет расстояние от каждого объекта до центров кластеров и приписывает объекты к ближайшему кластеру.

Шаг 5. Вновь вычисляются центры тяжести кластеров,

Шаг i. этот процесс повторяется до тех пор, пока центры тяжести не перестанут «мигрировать» в пространстве.

Перейдем на вкладку ***Advanced*** и выберем переменные для анализа. В диалоговом окно ***Select variables for analysis***, выделим все 27 параметров.

На поле ***Claster*** надо выбрать объекты для кластеризации. Так как цель исследования — кластеризация предприятий, относящиеся к наблюдениям, выберем ***Cases (rows)*** (наблюдения (строки)).

В поле ***Number of clasters*** (число кластеров) нужно определить число групп, на которые хотим разбить предприятия. Запишем в это поле число 5.

В поле ***Number of iterations*** (число итераций) задается максимальное число итераций, используемых при построении классов.

Группа опций ***Initial cluster centers*** позволяет задать начальные центры кластеров. Выберем *Choose observations to maximize initial between-cluster distances* (выбрать наблюдения, максимизирующие начальные расстояния между кластерами). Перейдем в окно результатов ***k-means Clustering Results*** (рис. 2).

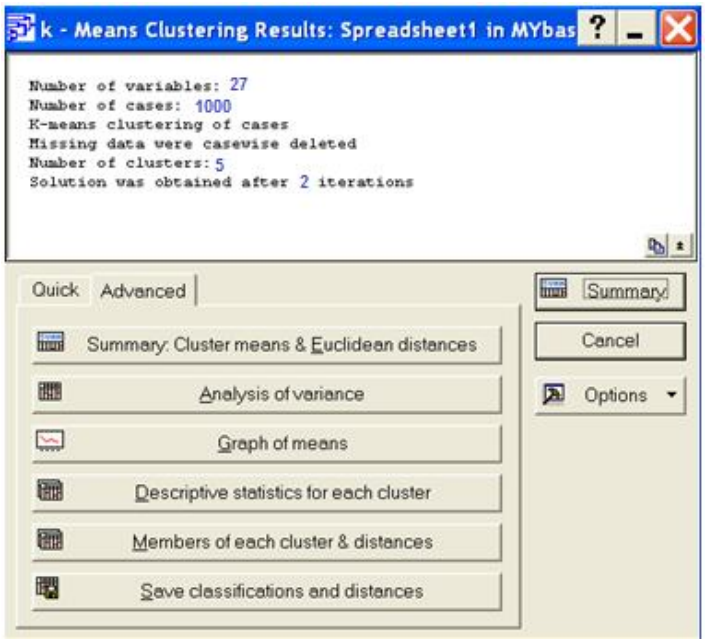

Рисунок 2. Окно результатов **k-means Clustering Results**

В верхней информационной части окна представлены следующие данные:

- *Number of variables (количество* переменных) 27;
- *Number of cases* (число наблюдений) 1000;
- *k-means clustering of cases* (метод *k-средних);*
- *Missing data were casewise deleted* (обработка пропущенных значений опущена);
- *Number of clusters* (число кластеров) 5;
- *Solution was obtained after* 2 *iterations* (решение было найдено после 2 итераций).

Откроем вкладку ***Advanced***, так как она содержит более подробную информацию о результатах анализа.

Выберем опцию ***Summary***: Cluster means & Euclidean distances предназначенную для вывода таблиц:

- Cluster means — указаны средние для каждого кластера (усреднение производится внутри кластера),
- Euclidean distances — евклидовы расстояния и квадраты евклидовых расстояний между кластерами. Рассмотрим таблицу, в которой указаны евклидовы расстояния и квадраты евклидовых расстояний между кластерами (рис.3).

Cluster Number	Euclidean Distances between Clusters Distances below diagonal Squared distances above diagonal				
	No. 1	No. 2	No. 3	No. 4	No. 5
No. 1	0,000000	0,537991	0,420104	1,017653	3,152336
No. 2	0,733479	0,000000	1,867715	2,929854	1,241899
No. 3	0,648154	1,366644	0,000000	0,141079	5,632123
No. 4	1,008788	1,711682	0,375605	0,000000	7,265060
No. 5	1,775482	1,114405	2,373209	2,695378	0,000000

Рисунок 3. Евклидовы расстояния и квадраты евклидовых расстояний

Евклидово расстояние – геометрическое расстояние в многомерном пространстве. В нашем случае это расстояние между наборами показателей (L1-A6) для каждого предприятия и оно эквивалентно расстоянию между предприятиями соответственно выбранным показателям. Чем меньше расстояние между объектами, тем они более схожи. Квадрат евклидова расстояния используют, если необходимо придать большие веса более отдаленным друг от друга объектам.

Вернемся к таблице значения под диагональю – есть евклидово расстояние, а над диагональю соответственно квадрат евклидова расстояния. Наибольшее расстояние между кластерами №4 и №5, то есть они менее всего схожи, что соответствует группам кредит и отказ. Почти на равных расстояниях между собой кластеры 3 и 1; 1 и 2 (высокий риск и средний риск; средний риск и низкий риск). Кластеры находятся на больших расстояниях друг от друга, так как евклидовы расстояния больше единицы.

Выберем опцию ***Analysis of variance***, программа выведет таблицу дисперсионного анализа (рис.4)

Variable	Analysis of Variance					
	Between SS	df	Within SS	df	F	signif. p
L1	97,71191	4	1,288094	95	1801,622	0,00
L3	97,50985	4	1,490150	95	1554,111	0,00
P1	97,42732	4	1,572678	95	1471,312	0,00
F1	97,68463	4	1,315369	95	1763,772	0,00
F2	97,46851	4	1,531487	95	1511,523	0,00
F3	97,88728	4	1,112717	95	2089,322	0,00
F4	97,68868	4	1,311317	95	1769,294	0,00
R1	97,80931	4	1,190689	95	1950,947	0,00
R2	96,44524	4	2,554764	95	896,590	0,00
R3	97,13944	4	1,860558	95	1239,984	0,00
R4	97,15035	4	1,849648	95	1247,438	0,00
R5	97,33225	4	1,667747	95	1386,086	0,00
A2	95,58330	4	3,416702	95	664,414	0,00
A4	94,03362	4	4,966385	95	449,683	0,00
A5	94,45660	4	4,543396	95	493,759	0,00
A6	95,84375	4	3,156250	95	721,201	0,00

Рисунок 4. Таблица дисперсионного анализа

В таблице приведены значения межгрупповых (Between SS) и внутригрупповых (Within SS) дисперсий признаков. Чем меньше значение внутригрупповой дисперсии и больше значение межгрупповой дисперсии, тем лучше признак характеризует принадлежность объектов к кластеру и тем «качественнее» наша кластеризация. Признаки с большими значениями *p* (например, больше 0,05) можно из процедуры кластеризации исключить. В нашем случае: для любого признака $p<0,05$, а значит никакой из рассматриваемых признаков исключать не будем.

По всем параметрам межгрупповая дисперсия больше 71, а внутригрупповая меньше 28. Лучше всего принадлежность объектов кластеру характеризуют показатели *F1, F4, A2 и A5,* так как они соответствуют наибольшей разнице между межгрупповой и внутригрупповой дисперсиями. Хуже всего (т.е. соответствуют наименьшей разнице дисперсий) характеризуют принадлежность – показатели *R1* и *R2*. Параметры F и p также характеризуют вклад признака в разделение объектов на группы. Лучшей кластеризации соответствуют большие значения первого и меньшие значения второго параметра. Из таблицы видно, что указанные выше наилучшие показатели соответствуют максимальной разнице *(F – p)*.

Graph of means позволяет просмотреть средние значения для каждого кластера на линейном графике (рис. 5).

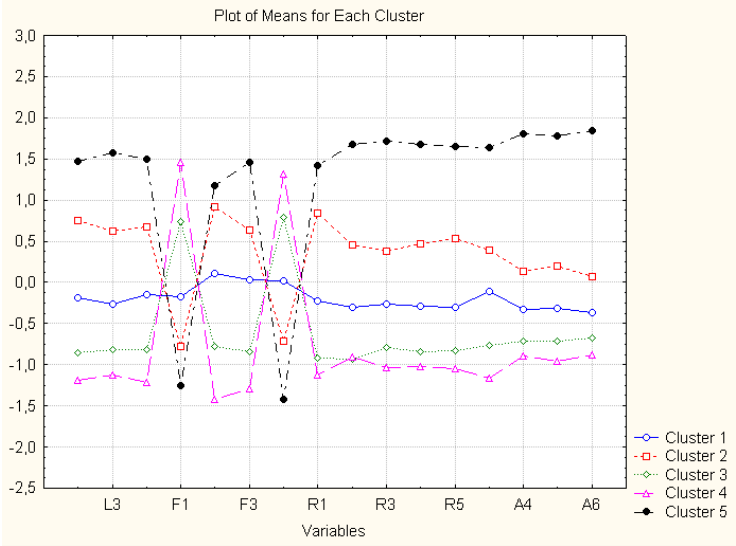

Рисунок 5. Линейный график кластеров

Наглядно прослеживаются 5 кластеров. Средние всех показателей значительно отличаются друг от друга. Это свидетельствует о качественном разбиении на группы. Как показывает график, расстояние между средними характеристик кластеров большое, также общее расстояние между центрами кластеров значительно, что свидетельствует об успешной кластеризации.

Сохраним результаты классификации в файле *STATISTICA* для дальнейшего исследования. Для этого выберем опцию ***Save classifications and distances.*** При этом в новом файле каждому наблюдению программой присваивается номер кластера, к которому он был отнесен при классификации. (рис.6)

Spreadsheet1 in MYbase.stw

	17 F3	18 F4	19 R1	20 R2	21 R3	22 R4	23 R5	24 A2	25 A4	26 A5	27 A6	28 CASE_NO	29 CLUSTER	30 DISTANCE
C_13	-1,286	1,364	-1,130	-0,949	-1,030	-1,019	-1,060	-1,129	-0,881	-0,943	-0,898	13	4	0,04
C_14	-1,302	1,304	-1,119	-1,091	-1,041	-1,017	-1,032	-1,176	-0,898	-0,963	-0,899	14	4	0,05
C_15	-1,258	1,257	-1,099	-0,984	-1,048	-1,038	-1,038	-1,225	-0,886	-0,959	-0,882	15	4	0,04
C_16	-1,331	1,369	-1,145	-0,874	-1,011	-1,013	-1,045	-1,152	-0,871	-0,930	-0,878	16	4	0,03
C_17	-1,281	1,310	-1,119	-0,839	-1,020	-1,026	-1,043	-1,163	-0,885	-0,978	-0,865	17	4	0,03
C_18	-1,299	1,381	-1,124	-0,916	-1,014	-1,024	-1,034	-1,155	-0,885	-0,932	-0,872	18	4	0,03
C_19	-1,323	1,214	-1,133	-0,900	-1,014	-1,028	-1,051	-1,189	-0,898	-0,975	-0,870	19	4	0,03
C_20	-1,230	1,390	-1,117	-1,061	-1,051	-1,025	-1,060	-1,089	-0,915	-0,940	-0,899	20	4	0,05
C_21	-0,837	0,843	-0,892	-0,996	-0,775	-0,834	-0,722	-0,697	-0,706	-0,672	-0,654	21	3	0,06
C_22	-1,042	0,733	-0,902	-0,905	-0,757	-0,726	-0,695	-0,884	-0,713	-0,750	-0,630	22	3	0,14
C_23	-0,902	0,782	-1,009	-0,911	-0,769	-0,851	-0,848	-0,740	-0,690	-0,696	-0,649	23	3	0,05
C_24	-0,902	0,839	-0,924	-0,992	-0,784	-0,839	-0,911	-0,676	-0,704	-0,742	-0,580	24	3	0,06
C_25	-0,769	0,863	-0,893	-1,015	-0,819	-0,794	-0,778	-0,827	-0,686	-0,672	-0,771	25	3	0,08
C_26	-0,810	0,782	-0,965	-1,002	-0,861	-0,798	-1,018	-0,738	-0,734	-0,621	-0,579	26	3	0,08
C_27	-0,672	1,030	-0,959	-0,850	-0,776	-0,880	-0,841	-0,786	-0,677	-0,692	-0,621	27	3	0,09
C_28	-0,887	0,646	-0,921	-0,775	-0,791	-0,823	-0,939	-0,832	-0,749	-0,751	-0,701	28	3	0,07
C_29	-0,846	0,746	-0,802	-1,077	-0,874	-0,766	-0,831	-0,778	-0,662	-0,793	-0,651	29	3	0,09
C_30	-0,674	0,984	-0,967	-0,775	-0,681	-0,849	-0,710	-0,751	-0,640	-0,734	-0,692	30	3	0,10
C_31	-0,887	0,893	-0,908	-0,994	-0,859	-0,901	-0,778	-0,728	-0,739	-0,679	-0,666	31	3	0,07
C_32	-0,768	0,362	-1,021	-1,120	-0,849	-0,855	-0,860	-0,755	-0,710	-0,695	-0,717	32	3	0,13
C_33	-0,911	0,851	-0,862	-0,956	-0,758	-0,830	-0,790	-0,750	-0,793	-0,833	-0,702	33	3	0,11
C_34	-0,793	0,961	-0,934	-0,683	-0,792	-0,853	-0,831	-0,780	-0,614	-0,715	-0,725	34	3	0,10
C_35	-0,795	0,857	-0,878	-1,159	-0,807	-0,744	-0,822	-0,817	-0,751	-0,683	-0,663	35	3	0,09

Рисунок 6. Часть таблицы с результатами кластерного анализа

Целю кластерного анализа являлась верификация полученной виртуальной базы российских предприятий. Полученные результаты полностью подтверждают проведенную классификацию:

- к кластеру под номером 4 относятся наблюдения (предприятия), которые относятся к классу «отказ»,
- к кластеру под номером 3 относятся наблюдения (предприятия), которые относятся к классу «высокий риск»,
- к кластеру под номером 1 - наблюдения (предприятия), которые относятся к классу «средний риск»,
- к кластеру под номером 2 – наблюдения, относящиеся к классу «низкий риск»
- под номером 5 – наблюдения класса «кредит».

Для удобной работы с переменными, принимающие текстовое значение, в программе STATISTICA реализован механизм двойной записи: каждому текстовому значению переменной ставится в соответствие некоторое число (и наоборот). Это соответствие может быть установлено самим пользователем. Воспользуемся данным механизмом двойной записи для осуществления большей наглядности результатов кластерного анализа.

Двойным щелчком на имени переменной (CLUSTER) зайдем в окно спецификации переменной. Нажмем кнопку Text Labels и перейдем к редактору текста ярлыков (рис. 7).

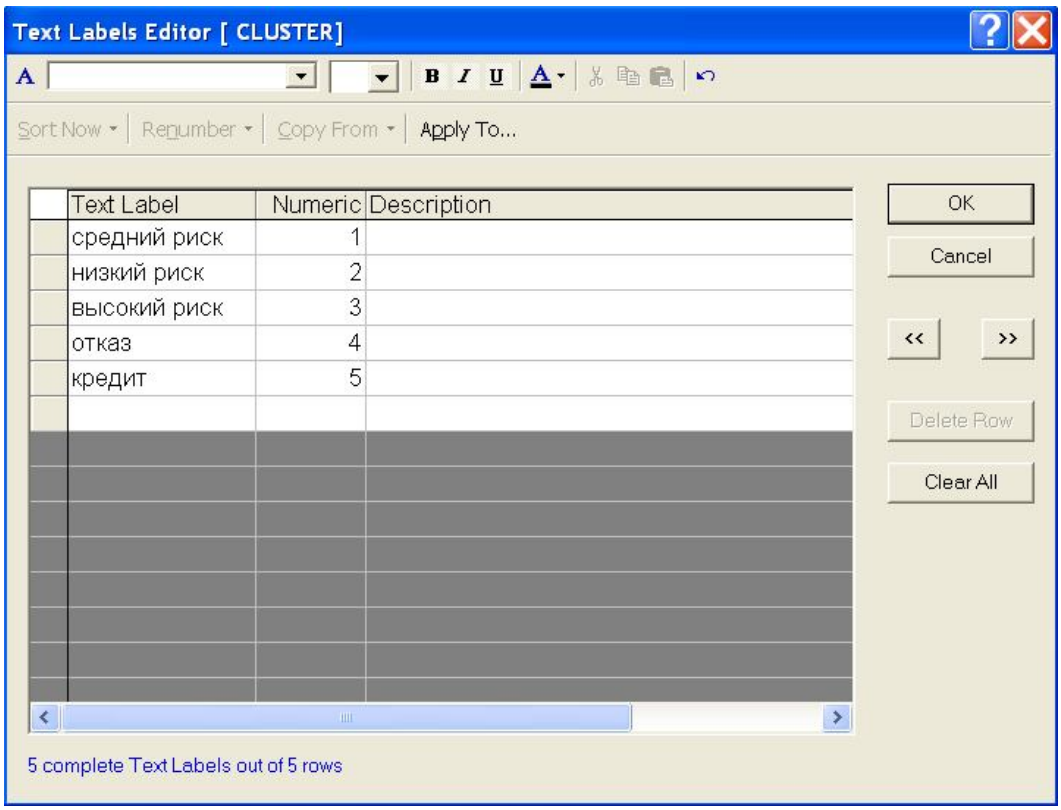

Рисунок 7. Редактор текста ярлыков

Каждый из пяти кластеров введем в поле Numeric, а соотвествующие текстовые названия полученных групп – в поле Text Label. Поле Description хранит краткое описание значений переменной и не является обязательным для заполнения. тогда для переключения отображения с числовых значений на текстовые и наоборот надо нажать на понели инструментов на кнопку Show/Hide Text Labels. Наличие такого механизма двойной записи значительно упрощает работу с текстовыми переменными.

Мы убедились в том, что данные о 1000 предприятиях качественно определяют принадлежность к классам и пригодны для дальнейшего исследования.