Б. Б. БАКШИЛОВ

КЛАСТЕРНЫЙ АНАЛИЗ БАНКА НА ОСНОВЕ ЕГО ФИЛИАЛОВ

Кластерный анализ решает задачу построения *классификации*, т.е. разделения исходного множества объектов на группы (классы, кластеры). При этом предполагается, что у исследователя нет исходных допущений ни о составе классов, ни об их отличии друг от друга. Приступая к кластерному анализу, исследователь располагает лишь информацией о характеристиках (признаках) для объектов, позволяющей судить о сходстве (различии) объектов, либо только данными об их попарном сходстве (различии). В литературе часто встречаются синонимы кластерного анализа: автоматическая классификация, таксономический анализ, анализ образов (без обучения).

Варианты кластерного анализа – это множество простых вычислительных процедур, используемых для классификации объектов. *Классификация* объектов – это группирование их в классы так, чтобы объекты в каждом классе были более похожи друг на друга, чем на объекты из других классов. Более точно, кластерный анализ – это процедура упорядочивания объектов в сравнительно однородные классы на основе попарного сравнения этих объектов по предварительно определенным и измеренным критериям.

Чаще всего кластерный анализ используется для выделения объектов, однородных по значениям набора признаков. Задачей наших исследований является использование кластерного

анализа для выявления объектов, однородных по значениям набора признаков, характеризующих основные показатели операционного риска по объектам (филиалам) одного из банков РК.

Задача классификации объектов по характеру сумм ущерба, их возмещения и остатка может быть представлена как поиск решения, позволяющего разделить множество объектов с определенным набором признаков на некоторое число групп сходных объектов по этим признакам. Следовательно, в результате кластерного анализа при помощи предварительно заданных переменных (признаков) формируются группы объектов (наблюдений). Члены одной группы (одного кластера) должны обладать схожими проявлениями переменных, а члены различных групп – различными.

Таким образом, перед нами стоит задача кластеризации N объектов, характеризуемых n признаками. Если подойти к решению этой задачи прямо, то найти оптимальную и объективную классификацию будет трудно либо невозможно.

Существует множество вариантов кластерного анализа, но наиболее широко используются методы, объединенные общим названием иерархический кластерный анализ (*Hierarchical Cluster Analysis*) [1–4]. В дальнейшем под кластерным анализом мы будем подразумевать именно эту группу методов.

Кластерный анализ объектов, для которых заданы значения количественных признаков, начинается с расчета различий для всех пар объектов. Пользователь может выбрать по своему усмотрению меру различия. В качестве меры различия выбирается расстояние между объектами в P-мерном пространстве признаков, чаще всего – евклидово расстояние или его квадрат. В данном случае $P = 2$ и евклидово расстояние между объектами i и j определяется формулой:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},$$

где x – это значения одного, а y – другого признака.

На первом шаге кластерного анализа путем перебора всех пар объектов определяется пара (или пары) наиболее близких объектов, которые объединяются в первичные кластеры. Далее на каждом шаге к каждому первичному кластеру присоединяется объект (кластер), который ближе к нему. Этот процесс повторяется до тех пор, пока все объекты не будут объединены в один кластер. Для N объектов эта процедура требует N-1 стадий объединения.

Изложенный метод проиллюстрируем на примере 3-х показателей риска (переменные) по 20-ти филиалам банка за последние 12 месяцев (август 2007-август 2008гг.) (табл. 1).

Таблица 1

№ п/п	Сумма ущерба	Сумма возмещения	Остаток ущерба
1	1 106,1	681,0	425,1
2	936,3	147,9	788,4
3	34 151,8	29 069,2	5 082,6
4	9 586,0	6 547,4	3 038,6
5	20 477,2	122,8	20 354,4
6	4,5	4,5	0,0
7	2 354,9	1 504,6	850,3
8	8 720,3	274,6	8 445,7
9	43 323,4	43 030,5	292,9
10	31,5	0,0	31,5
11	1 466,4	845,8	620,6
12	5 012,7	116,5	4 896,2
13	2 612,6	172,0	2 440,6
14	5 899,1	815,9	5 083,2
15	552,0	43,4	508,6
16	2 413,8	2 266,8	147,0
17	677,7	32,5	645,2
18	3 280,8	51,1	3 229,7
19	408,2	212,7	195,5
20	7 390,2	475,0	6 915,2

Основным результатом применения иерархического кластерного анализа является *дендрограмма* – графическое изображение последовательности объединения объектов в кластеры.

Пакет *SPSS* [5] предлагает, в общей сложности, 7 различных методов объединения.

В табл. 2 приведен обзор принадлежности, из которого можно выяснить очередность построении кластеров, а также их оптимальное количество.

По двум колонкам, расположенным под общей шапкой – «Объединение в кластеры», можно увидеть, что на первом шаге были объединены объекты 6 и 10, эти два объекта максимально похожи друг на друга и отдалены на очень малое расстояние. Два данных наблюдения образовывают кластер с номером 6, в то же время, как объект 10 в обзорной таблице больше не появляется. На следующем шаге происходит объединение объектов 15 и 17, затем 1 и 11 и т.д.

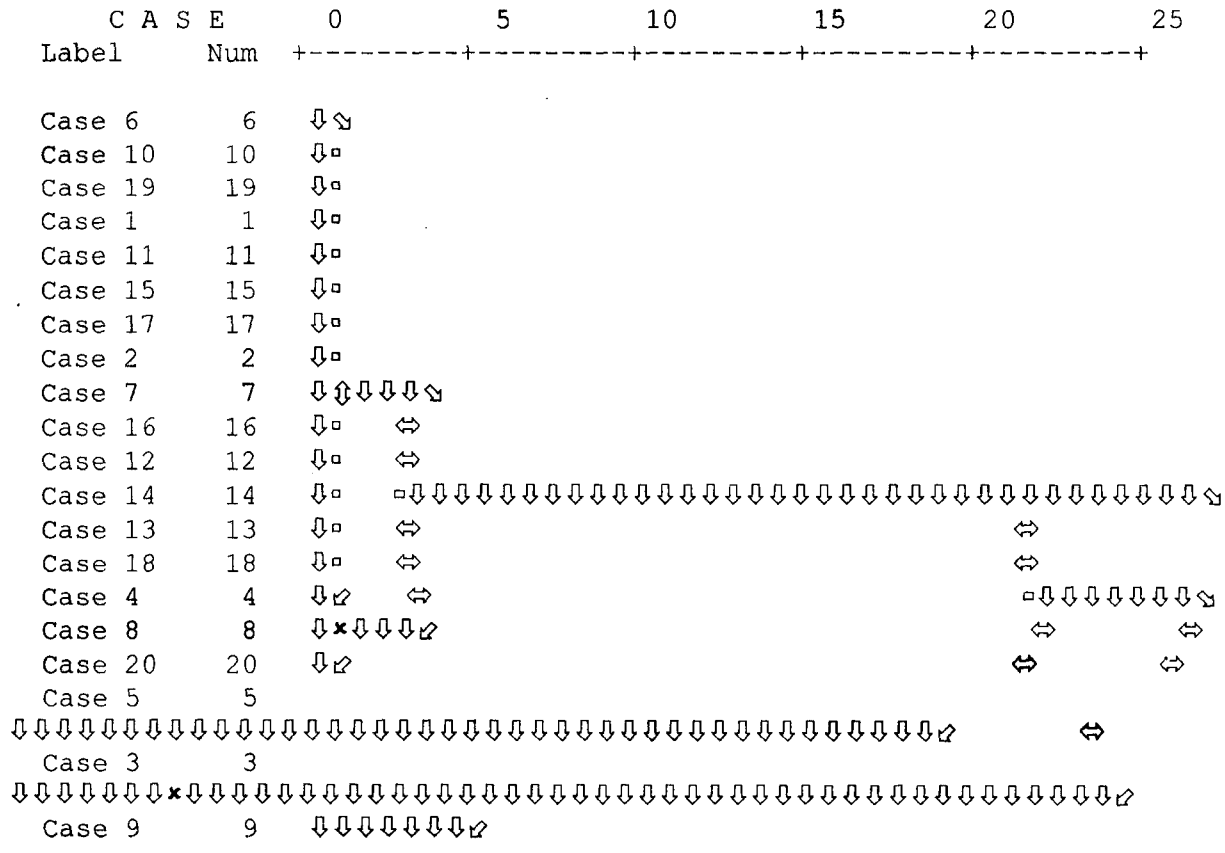
Для определения, какое количество кластеров следовало бы считать оптимальным, решающее значение имеет показатель, выводимый под

Таблица 2. **Порядок агломерации**

Шаг	Объединение в кластеры		Коэффициенты
	Кластер 1	Кластер 2	
1	6	10	4,896E-05
2	15	17	9,337E-04
3	1	11	2,827E-03
4	6	19	2,878E-03
5	2	15	3,036E-03
6	1	2	8,245E-03
7	12	14	1,108E-02
8	1	6	2,018E-02
9	7	16	2,639E-02
10	13	18	3,066E-02
11	1	7	5,589E-02
12	8	20	0,116
13	12	13	0,260
14	4	12	0,623
15	1	4	0,778
16	1	8	2,116
17	3	9	3,170
18	1	5	16,937
19	1	3	20,599

заголовком «Коэффициенты». Под этим коэффициентом подразумевается расстояние между двумя кластерами, определенное на основании выбранной дистанционной меры с учетом предусмотренного преобразования значений. На этапе, где мера расстояния между двумя кластерами увеличивается скачкообразно, процесс объединения в новые кластеры необходимо остановить, так как в противном случае произошло бы объединение кластеров, находящихся на относительно большом расстоянии друг от друга.

В нашем случае – это скачек после 17-го шага. Это означает, что для данных, включающих 20 объектов (наблюдений), оптимальным является решение с тремя кластерами. Следовательно, оптимальным считается число кластеров, равное разности количества наблюдений (здесь 20) и количества шагов, после которого коэффициент увеличивается скачкообразно (здесь 17). После определения оптимального количества кластеров организуем для каждого объекта вывод информации о принадлежности к кластеру (табл. 3).



Дендрограмма

Таблица 3. **Принадлежность к кластеру**

№ п/п	3 кластера
1	1
2	1
3	2
4	1
5	3
6	1
7	1
8	1
9	2
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1

Из табл. 3 видно, что в первый кластер входят 17 объектов, во второй – два и в третий – 1 объект.

В заключение приведем дендрограмму (рис.), которая визиализирует процесс слияния, приведенный в обзорной таблице 2 порядка агломерации. Она идентифицирует объединенные кластеры и значения коэффициентов на каждом шаге. При этом отображаются не исходные значения коэффициентов, а значения, приведенные к шкале от 0 до 25.

Таким образом, проведенный кластерный анализ позволил распределить всю совокупность объектов по трем кластерам, объединив в них объекты, имеющие схожую структуру. Выделенные в результате проведенного кластерного анализа качественно однородные группы объектов, составляющие исследуемую совокупность, используются в последующем для проведения соответствующего анализа.

Кластерный анализ – это комбинаторная процедура, имеющая простой и наглядный результат. Широта возможного применения кластерного анализа очевидна настолько же, насколько очевиден и его смысл. Классифицирование или разделение исходного множества объектов на различающиеся группы – всегда первый шаг в любой умственной деятельности, предваряющий поиск причин обнаруженных различий.

ЛИТЕРАТУРА

1. *Большаков А.А., Каримов Р.Н.* Методы обработки многомерных данных и временных рядов. М.: Горячая линия – Телеком, 2007. 522 с.

2. *Вуколов Э.А.* Основы статистического анализа. М.: ИНФРА-М, 2004. 464 с.

3. *Новорожкина Л.И., Арженовский С.В.* Многомерные статистические методы в экономике. М.: Дашков и К°, 2007. 224 с.

4. *Орехов Н.А., Левин А.Г., Горбунов Е.А.* Математические методы и модели в экономике. М.: ЮНИТИ-ДАНА, 2004. 302 с.

5. *Плис А.И., Сливина Н.А.* Практикум по прикладной статистике в среде SPSS. М.: Финансы и статистика, 2004. 288 с.

Резюме

Зерттеу барысында кластерлік анализ мәселелерінің теориялық аспектілері қарастырылған. Қазақстан Республикасының екінші деңгейлі банктері бірінің кластеріне оның филиалдары негізінде ғылыми іздестірулер жүргізілген, осы ретте бір филиалдың екіншісіне қатыстылық қағидасы сатылы түрде түсінідіріліп, филиалдардың агломерация тәртібі беріледі.

Жүргізілген зерттеулердің нәтижесінде құрамында 20 (жиырма) филиалы бар объектілер жиынтығы 3 кластерге бөлінді. Кластерлік талдау негізінде дендрограмма жасалды, ол дендрограммада аталған кластерлерді көрнекті түрде көруге болады.

Болашақтағы зерттеулер белгілі кластерлерде пайда болуы мүмкін опрециондық тәуекелдер ықтималдығын болжамдай алатын математикалық модельдерді құруға мүмкіндік береді.

Summary

In scientific work theoretical aspects of Hierarchical Cluster Analysis have been considered. The author researches the cluster analysis one of the second level banks of the Republic of Kazakhstan on the basis of its branches, where the principle of an accessory of one branch to another are explained step by step, the order of agglomeration of branches is given.

As a result of the scientific work all objects in quantity of 20 (twenty) branches have been divided on 3 clusters. On the basis of cluster analysis a dendrogramma has been deduced which allows a visual presence of these clusters.

The further research will allow to construct mathematical models which will predict probability of occurrence of operational risks in defined clusters.

КазЭУ им. Т. Рыскулова *Поступила 10.06.08г.*