УДК 33:519.2

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МНОГОМЕРНОЙ СТАТИСТИКИ ДЛЯ АНАЛИЗА СОЦИАЛЬНОЙ И ЭКОНОМИЧЕСКОЙ ИНФОРМАЦИИ

Фомина Е.Е.

Тверской государственный технический университет

Аннотация. Анализ социально-экономической информации, которая описывается в большинстве случаев большим набором различных переменных, достаточно трудоемкая задача. Использование методов математической статистики и программного обеспечения позволит существенное упростить этот процесс и получить результаты, на основе которых могут быть приняты значимые решения. В статье рассматривается методика совместного использования кластерного анализа и метода дерева классификации для структурирования и классификации социально-экономической информации. Применение методики продемонстрировано на примере обработки информации по уровню жизни в регионах РФ.

Ключевые слова: анализ экономической информации, кластерный анализ, деревья классификации

1. Введение

Социально-экономические процессы описываются, как правило, большим набором переменных, совместный анализ которых достаточно трудоёмкая задача. В свою очередь, изучение взаимосвязи между этими переменными позволит адекватно оценить состояние развития процесса или объекта, структурировать информацию, выявить наиболее значимые переменные и, как следствие, выработать стратегическое решение. Эффективным инструментом исследования, в этом случае, может оказаться применение математических методов с использованием вычислительной техники и прикладного программного обеспечения.

К таким методам можно отнести корреляционно-регрессионный анализ, кластерный и дискриминантный анализ, факторный анализ [11, 12], нейронные сети, оптимизационные модели, модели межотраслевого баланса, модели теории массового обслуживания и другие [6].

В настоящей статье рассматривается методика совместного применения кластерного анализа и метода дерева классификации для исследования и структурирования социально-экономической информации, поиска наиболее значимых переменных, обнаружения скрытых закономерностей.

Применение методики продемонстрировано на примере обработки данных, содержащих информацию по уровню жизни населения в регионах Российской Федерации.

2. Методы исследования

Как было сказано выше, методика включает в себя применение методов кластерного анализа (КА) и дерева классификации (ДК).

Под кластерным анализом понимается группа математических методов, предназначенных для формирования групп «близких» между собой объектов, описываемых некоторой системой признаков, по информации о расстояниях или связях между ними [5, 7].

Основное назначение КА – разбиение выборки на однородные группы или кластеры, что приводит к сокращению объема информации. Упрощается исследование структуры данных и их дальнейшая обработка. Имеется возможность обнаружения новизны, т.е. редких объектов, которые не удаётся отнести ни к одному из кластеров [5, 3].

Алгоритм выполнения КА включает в себя следующие этапы:

Корректный выбор переменных, на основании которых будет проведена процедура разбиения на кластеры. Значимость этого этапа не вызывает сомнений, так как включение посторонних переменных может существенно исказить результат.

Стандартизация переменных, позволяющая привести их значения к единому диапазону.

Выбор меры сходства. Основная задача КА – объединение объектов в кластеры. Для оценки близости объектов и последующего их объединения используют меры сходства, которые делятся на две группы: меры сходства типа расстояния: евклидово расстояние, квадрат евклидова расстояния, манхэттенское расстояние, расстояние Чебышева, процент несогласия, степенное расстояние; мера сходства типа корреляции: расстояние Пирсона [3, 4, 8].

Выбор подходящего для решаемой задачи метода кластеризации. Методы кластеризации делятся на две группы: иерархические; неиерархические [3].

Группа иерархических методов кластеризация предполагает построение дендрограммы или древовидной диаграммы, которая содержит шаги последовательного объединения объектов в кластеры [3, 4].

Группа неиерархических методов (методов *k*-средних) предполагает задание числа кластеров *k*, на которые будет разделено все множество объектов [3, 4].

Определение оптимального количества кластеров. На этом этапе следует руководствоваться следующими соображениями: число кластеров должно быть таким, чтобы явно прослеживалась наглядность данных; относительные размеры кластеров должны быть достаточно выразительными.

Оценка устойчивости полученного решения.

Содержательная интерпретация кластеров исходя из кластерных центроидов, результатов дисперсионного анализа и графика средних.

Существенным преимуществом метода КА является возможность его применения для исследования совокупности объектов, описываемых переменными любого типа. Применение метода позволяет выделить в исходном массиве информации однородные группы, однако кластерный анализ не дает возможности сформулировать правило, которое позволит проводить классификацию объектов. Для решения этой задачи может использоваться метод дискриминантного анализа. Однако он имеет ряд ограничений, связанных с метризуемостью пространства исходных данных.

Альтернативой может выступать метод деревьев классификации, позволяющий изучать статистическую взаимосвязь между одной зависимой и группой независимых переменных, а также определять принадлежность объектов к тому или иному классу в зависимости от значений переменных, характеризующих объекты [2, 10].

Метод не накладывает каких-либо ограничений на тип исходных данных, позволяет визуализировать результаты классификации, а также формулировать набор классификационных плавил, что существенно упрощает интерпретацию.

ДК характеризуются построением дерева, состоящего из корневого узла (представляющего собой всю выборку), дочерних и родительских узлов, а также терминальных узлов, т.е. окончательных узлов, которые далее не разбиваются. Каждой вершине ставится в соответствие правило, согласно которому объекты относятся к тому или иному классу.

На сегодняшний день разработано большое число алгоритмов, позволяющих реализовывать деревья классификаций. Рассмотрим самые популярные из них [2, 1]:

алгоритм CART предназначен для построения бинарного дерева решений, т.е. такого дерева, в котором каждый узел при разбиении имеет только двух потомков. Основу алгоритма составляет проверка оценочной функции, которая базируется на идее уменьшения неопределенности в узле. Метод может использоваться для анализа как количественных, так и категориальных переменных;

алгоритм C4.5 предназначен для многомерного расщепления. Метод используется для анализа как количественных, так и категориальных независимых переменных, но зависимая переменная должна быть категориальной.

Методика комбинация КА и ДК позволит выработать эффективный алгоритм структурирования и анализа информации, который включает в себя два этапа: разбиение множества объектов на кластеры и применение ДК для построения решающего правила распределения объектов.

3. Материалы исследования

Применение методики продемонстрировано на примере. В качестве материалов исследования выступала база данных (включающая 83 региона), содержащая информацию по остатку денежных средств семьи после минимальных расходов для семей с двумя и тремя детьми по регионам РФ [9]. Фрагмент базы данных представлен в табл. 1.

Таблица 1 - Фрагмент таблицы базы данных по остатку денежных средств

№	Регион	Остаток денежных средств семьи после минимальных расходов (руб. в месяц), в 2013 году	
		2 детей (переменная 1)	3 детей (переменная 2)
1	Ямало-Ненецкий автономный округ	90 545	78 912
2	Чукотский автономный округ	86 160	72 742
3	Москва	69 952	60 611
…	…	…	…

Уровень жизни в регионе оценивается по остатку денежных средств, который определяется как разница между доходами двух взрослых родителей (две средние заработные платы по региону) и четырьмя (или пяти, в зависимости от количества детей) прожиточными минимумами. Необходимо было разбить исходную выборку регионов на группы (кластеры) и сформулировать наглядное правило для классификации регионов. Обработка данных проводилась в пакете STATISTICA.

Первый этап исследование – применение процедуры кластерного анализа для разбиения объектов на группы. Использовался алгоритм метода *к*-средних, число кластеров полагалось равным трем. Результаты кластерного анализа позволили выделить три однородные группы с высоким (8 регионов), низким (53 региона) и средним (22 региона) уровнем остатка денежных средств (табл. 2, рис. 1). Как показывает анализ графика средних, регионы первого кластера значительно удалены по значениям показателей от регионов 2 и 3 кластеров. Данные дисперсионного анализа говорят о том, что обе переменные являются значимыми (p значение менее 0,05) (табл. 3).

Таблица 2 - Характеристики кластеров

Кластер	Переменная	Среднее	Стандартное отклонение	Коэффициент вариации
1	2 детей	66399,25	15386,18	23%
	3 детей	54943,00	14724,60	27%
2	2 детей	29116,64	6880,989	24%
	3 детей	19800,27	6961,364	35%
3	2 детей	15151,92	4063,754	27%
	3 детей	8450,91	3960,653	47%

Таблица 3 - Дисперсионный анализ

Переменная	Между SS	Внутри SS	Значим. P
2 детей	1,9E+10	3,5E+09	4,4E-33
3 детей	1,5E+10	3,4E+09	1,1E-30

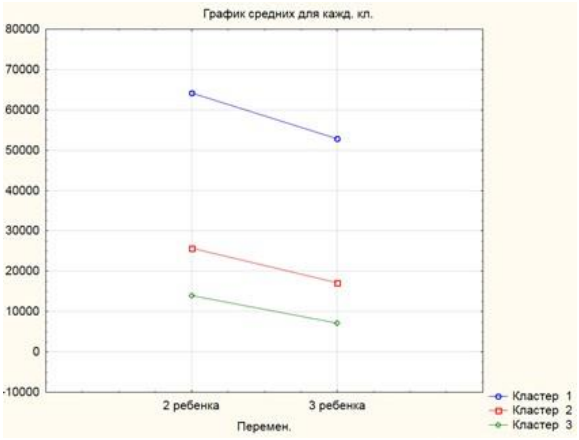

Рис. 1. Кластерный анализ: график средних

Первый кластер характеризуется высокими остатками денежных средств, что согласно методике говорит о высоком уровне жизни населения. Кластер составляют 8 регионов, к которым, в том числе, относится Москва, а также регионы Северо-Западного, Уральского и Дальневосточного федеральных округов (рис. 2).

Второй кластер характеризуется средними остатками денежных средств. В кластер входит 22 региона преимущественно Сибирского, Северо-Западного и Уральского федеральных округов (рис. 2).

Третий кластер, самый многочисленный, характеризуется невысокими остатками денежных средств. В него входит 53 региона преимущественно Центрального и Приволжского федеральных округов (рис. 2).

1 КЛАСТЕР	
1	Ямало-Ненецкий автономный округ
2	Чукотский автономный округ
3	Москва
4	Ханты-Мансийский автономный округ - Югра
5	Ненецкий автономный округ
6	Магаданская область
7	Сахалинская область
8	Санкт-Петербург

2 КЛАСТЕР	
1	Республика Саха (Якутия)
2	Московская область
3	Камчатский край
4	Мурманская область
5	Республика Коми
6	Ленинградская область
7	Тюменская область
8	Красноярский край
9	Иркутская область
10	Томская область
11	Республика Татарстан
12	Хабаровский край
13	Свердловская область
14	Забайкальский край
15	Республика Хакасия
16	Калужская область
17	Омская область
18	Республика Бурятия
19	Калининградская область
20	Республика Карелия
21	Кемеровская область
22	Челябинская область

3 КЛАСТЕР							
1	Амурская область	19	Оренбургская область	37	Владимирская область
2	Архангельская область	20	Удмуртская Республика	38	Карачаево-Черкесская Республика
3	Нижегородская область	21	Тамбовская область	39	Республика Марий Эл
4	Приморский край	22	Самарская область	40	Орловская область
5	Краснодарский край	23	Курская область	41	Ульяновская область
6	Новосибирская область	24	Рязанская область	42	Брянская область
7	Пермский край	25	Пензенская область	43	Курганская область
8	Вологодская область	26	Тверская область	44	Кировская область
9	Воронежская область	27	Саратовская область	45	Республика Мордовия
10	Белгородская область	28	Еврейская автономная область	46	Костромская область
11	Астраханская область	29	Ставропольский край	47	Псковская область
12	Тульская область	30	Волгоградская область	48	Кабардино-Балкарская Республика
13	Республика Тыва	31	Ростовская область	49	Смоленская область
14	Липецкая область	32	Чувашская Республика	50	Ивановская область
15	Республика Башкортостан	33	Республика Алтай	51	Алтайский край
16	Республика Ингушетия	34	Республика Адыгея	52	Республика Калмыкия
17	Новгородская область	35	Чеченская Республика	53	Республика Дагестан
18	Ярославская область	36	Республика Северная Осетия - Алания		

Рис. 2. Кластеры

Второй этап исследования включал применение к выделенным кластерам метода дерева

классификации для построения правила отнесения объектов к тому или иному кластеру (рис. 2).

Для реализации ДК использовался метод CART. В результате применения метода получено следующее дерево (рис. 3).

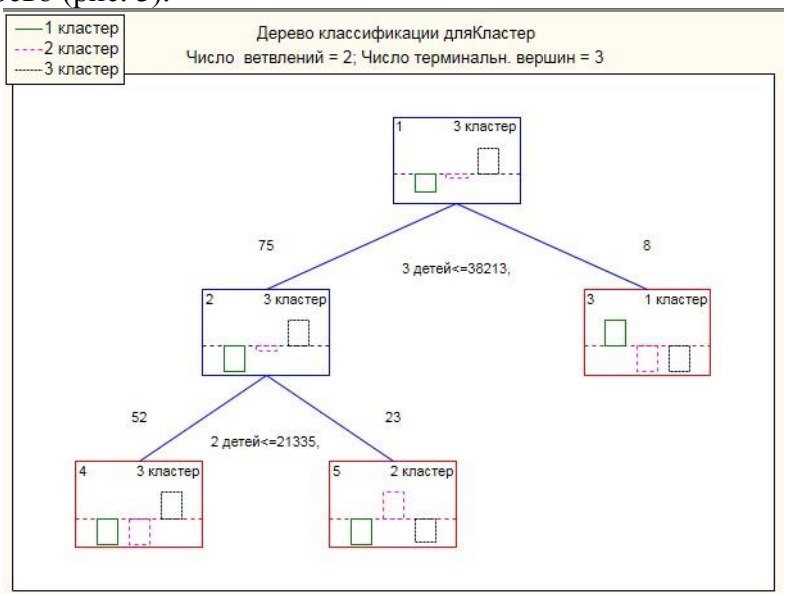

Рис. 3. Дерево классификации

Прокомментируем полученное дерево, начиная с корневого узла (1 вершина). Исходное количество регионов (83 региона) делится на две части: 75 объектов – третий кластер (2 вершина) и 8 объектов – 1 кластер (3 вершина) по следующему правилу: если остаток денежных средств семьи, состоящей из 3 детей менее или равен 38 213 руб., то регион принадлежит к третьему кластеру, иначе к первому. Далее происходит следующий этап анализа.

Для левой ветви: если остаток денежных средств семьи, состоящей из 2 детей менее или равен 21 335 руб., то регион принадлежит к третьему кластеру (4 вершина, включает 52 объекта), в противном случае – ко второму кластеру (5 вершина, включает 23 объекта).

Таким образом, можно построить следующее итоговое классификационное правило (рис. 4).

Рис. 4. Классификация объектов по кластерам

Результат классификации объектов составил 98,7%. Обе переменные оказались значимыми с рангами 100.

4. Выводы

Предложена методика совместного использования методов кластерного анализа и дерева классификации для структурирования и поиска статистических закономерностей в массиве социально-экономической информации.

Применение методики продемонстрировано на примере обработки данных по уровню жизни в регионах РФ. На первом этапе исследования был применен кластерный анализ, позволивший выделить и охарактеризовать три кластера, включающих регионы с высоким, средним и

низким уровнем остатка денежных средств. На следующем этапе, применяя дерево классификации, получены достаточно простые и наглядные решающие правила, согласно которым можно обоснованно отнести регион к тому или иному кластеру.

Список литературы

1. Анализ статистических данных с использованием деревьев решений. Режим доступа: http://math.nsc.ru/AP/datamine/decisiontree.htm (дата обращения 22.04.2018).

2. Бова А. Деревья решений как техника добычи данных // Социология: теория, методы, маркетинг. 2002. №1. С. 128-136.

3. Буреева Н.Н. Многомерный статистический анализ с использованием ППП «STATISTICA»: учебно-методический материал по программе повышения квалификации «Применение программных средств в научных исследованиях и преподавании математики и механики». Нижний Новгород: Нижегородский государственный университет им. Н.И. Лобачевского. 2007. 112 с.

4. Гайдышев И.П. Программное обеспечение анализа данных AtteStat. Руководство пользователя. Версия 13. 2012. 505 с.

5. Дюран Б. Кластерный анализ / Б. Дюран, П. Одел. М.: Статистика. 1977. 128 с.

6. Звягин Л. С. Актуальные экономико-математические методы исследования современных экономических процессов // Вопросы экономики и управления. 2015. №2. С. 1-6.

7. Королёв, М.А. Статистический словарь / гл. ред. М.А. Королёв. М.: Финансы и статистика. 1989. 623 с.

8. Меры расстояния и сходства между объектами. Режим доступа: https://studfiles.net/preview/1582407/page:3/ (дата обращения 18.12.2017).

9. Рейтинг регионов по уровню жизни семей – 2014. Режим доступа: http://riarating.ru/infografika/20140707/610622475.html (дата обращения 05.05.2018).

10. Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. М.: Научный мир. 2000. 352 с.

11. Фомина Е.Е. Факторный анализ и категориальный метод главных компонент: сравнительный анализ и практическое применение для обработки результатов анкетирования // Гуманитарный вестник. 2017. № 10 (60). С. 3.

12. Фомина Е.Е., Жиганов Н.К. Методика обработки результатов анкетирования с использованием методов многомерной и параметрической статистики // Вестник Пермского национального исследовательского политехнического университета. Социально-экономические науки. 2017. № 1. С. 106-115.

Сведения об авторах

Фомина Елена Евгеньевна - кандидат технических наук, доцент кафедры Информатики и прикладной математики Тверского государственного технического университета, E-mail: f-elena2008@yandex.ru

UDK 33:519.2

THE APPLICATION OF MULTIVARIATE STATISTICS METHODS FOR ANALYZE SOCIAL AND ECONOMIC INFORMATION

Fomina E.E.

Tver State Technical University

Abstract. The analysis of socio-economic information which is described by a large set of different variables is difficult task for researcher. The use of mathematical methods and specialized software will significantly simplify this process, classified the information, identify statistical patterns and get the results on the basis of which can be made important decisions. The article is devoted to the method of joint use the cluster analysis and the method of classification tree for structuring and classification of socio-economic information. The application of the method is demonstrated by the example of information processing on the level of life in the regions of Russian Federation.

Keywords: the analysis of economic information, cluster analysis, classification trees

References

1. Analiz statisticheskih dannyh s ispol'zovaniem derev'ev reshenij. Rezhim dostupa: http://math.nsc.ru/AP/datamine/decisiontree.htm (data obrashcheniya 22.04.2018).

2. Bova A. Derev'ya reshenij kak tekhnika dobychi dannyh // Sociologiya: teoriya, metody, marketing. 2002. №1. S. 128-136.

3. Bureeva N.N. Mnogomernyj statisticheskij analiz s ispol'zovaniem PPP «STATISTICA»: uchebno-metodicheskij material po programme povysheniya kvalifikacii «Primenenie programmnyh sredstv v nauchnyh issledovaniyah i prepodavanii matematiki i mekhaniki». Nizhnij Novgorod: Nizhegorodskij gosudarstvennyj universitet im. N.I. Lobachevskogo. 2007. 112 s.

4. Gajdyshev I.P. Programmnoe obespechenie analiza dannyh AtteStat. Rukovodstvo pol'zovatelya. Versiya 13. 2012. 505 s.

5. Dyuran B. Klasternyj analiz / B. Dyuran, P. Odel. M.: Statistika. 1977. 128 s.

6. Zvyagin L. S. Aktual'nye ehkonomiko-matematicheskie metody issledovaniya sovremen-nyh ehkonomicheskih processov // Voprosy ehkonomiki i upravleniya. 2015. №2. S. 1-6.

7. Korolyov, M.A. Statisticheskij slovar' / gl. red. M.A. Korolyov. M.: Finansy i stati-stika. 1989. 623 s.

8. Mery rasstoyaniya i skhodstva mezhdu ob"ektami. Rezhim dostupa: https://studfiles.net/preview/1582407/page:3/ (data obrashcheniya 18.12.2017).

9. Rejting regionov po urovnyu zhizni semej – 2014. Rezhim dostupa: http://riarating.ru/infografika/20140707/610622475.html (data obrashcheniya 05.05.2018).

10. Tolstova YU.N. Analiz sociologicheskih dannyh. Metodologiya, deskriptivnaya stati-stika, izuchenie svyazej mezhdu nominal'nymi priznakami. M.: Nauchnyj mir. 2000. 352 s.

11. Fomina E.E. Faktornyj analiz i kategorial'nyj metod glavnyh komponent: sravnitel'nyj analiz i prakticheskoe primenenie dlya obrabotki rezul'tatov anketirovaniya // Gumanitarnyj vestnik. 2017. № 10 (60). S. 3.

12. Fomina E.E., ZHiganov N.K. Metodika obrabotki rezul'tatov anketirovaniya s ispol'zovaniem metodov mnogomernoj i parametricheskoj statistiki // Vestnik Permskogo nacional'nogo issledovatel'skogo politekhnicheskogo universiteta. Social'no-ehkonomicheskie nauki. 2017. № 1. S. 106-115.

Author`s information

Fomina Elena Evgenievna - Candidate of technical Sciences, associate Professor of Informatics and applied mathematics, Tver State Technical University, E-mail: f-elena2008@yandex.ru