Ирина Гелиевна ЗАХАРОВА — заведующая кафедрой программного обеспечения математического факультета, доцент,
Дарья Алексеевна ПЛЕСОВСКИХ — студентка 5 курса

УДК 519.2, 681.3

ОБ ОДНОМ ПОДХОДЕ К КОМПЬЮТЕРНОЙ РЕАЛИЗАЦИИ ИЕРАРХИЧЕСКОГО КЛАСТЕРНОГО АНАЛИЗА

АННОТАЦИЯ. Предложена компьютерная реализация иерархического кластерного анализа. Рассмотрена математическая модель процесса классификации. Приведены алгоритмы построения классов. Приведены результаты вычислительного эксперимента на конкретном примере.

A computer realization of hierarchical cluster analysis is suggested. A mathematical model of the classification is considered. Algorithms of classes' construction are given. Results of the computer experiment based on some examples are also presented.

Настоящая работа посвящена созданию программного обеспечения, предназначенного для решения весьма широкого круга задач в самых разных областях — от медицины до управления. В частности, реализованный в нем метод позволяет моделировать те или иные ситуации на основе решения задач классификации и распознавания объектов (а в дальнейшем, и прогнозирования).

Итак, пусть задана совокупность объектов, каждый из которых описывается p признаками (исходная выборка). В пространстве описания каждый объект представляется p-мерным вектором или точкой, координаты которой есть значения соответствующих признаков. Исходную выборку требуется разбить на однородные в некотором смысле классы, причем априорная информация о числе групп, характере распределения объектов внутри каждой группы, обучающих выборках может и отсутствовать.

Полученные в результате разбиения классы называют кластерами, а методы их нахождения — кластерным анализом. Пользуясь методами кластерного анализа для решения конкретной задачи, исследователь пытается определить естественное расслоение объектов исходной выборки на более или менее ярко выраженные кластеры, лежащие на некотором расстоянии друг от друга в многомерном пространстве описания. В принципе, полученное решение может не представлять практического интереса, например, если вся выборка объединится в один кластер.

Итак, методы кластерного анализа позволяют представить изучаемую совокупность объектов как иерархическую систему кластеров — групп «схожих» объектов. При этом мера «схожести» на каждом уровне иерархии, естественно, различна.

Методы кластеризации довольно разнообразны. В них по-разному выбирается способ определения близости между кластерами и самими исходными объектами, используются различные алгоритмы вычислений [1, 3].

Очень существенным здесь является то, что результаты кластеризации подчас зависят от выбранного метода, и эта зависимость тем сильнее, чем менее явно изучаемая совокупность разделяется на характерные группы объектов. Поэтому результаты кластеризации могут быть дискуссионными. В этом случае рекомендуется использовать их лишь как предпосылку для содержательного анализа. Иногда полученные результаты можно обосновать, используя методику дискриминантого анализа.

В ряде специализированных пакетов имеется возможность реализации кластерного анализа данных [1]. Но, говоря о преимуществах применения специализированных статистических программных комплексов, нельзя забывать о том, что их использование исследователем, недостаточно компетентным в области статистических методов, — процесс весьма трудоемкий. (Попробуйте это сделать, взяв, например, пакет SPSS). Такая некомпетентность играет свою роль как при выборе метода обработки данных, так и на этапе интерпретации полученных результатов. Предварительный автоматический анализ исходных данных и построение решения в ходе непрерывного диалога с пользователем позволяет в определенной мере решить эти проблемы.

В настоящей работе предлагается компьютерная реализация иерархического кластерного анализа, делающая весь процесс прозрачным для пользователя. Как на стадии подготовки исходных данных, так и при построении системы кластеров реализована возможность анализа промежуточных результатов и коррекции процесса. Программный продукт представляет собой приложение Windows, файлы исходных данных могут иметь формат MS Excel (*. xls), MS Access (*. mdb), FoxPro (*. dbf).

Применяющийся метод основан на использовании агломеративной иерархической процедуры кластеризации. Такая процедура представляет собой пошаговый алгоритм, на каждом шаге которого происходит объединение множества объектов, подлежащих классификации, в непересекающиеся кластеры (число объектов, объединяемых в кластер, может быть произвольным). При этом каждое последующее объединение применяется к кластерам, полученным на предыдущем шаге.

В данной работе в качестве основы для построения процедуры применен следующий алгоритм. Задается монотонно возрастающая последовательность пороговых значений $\{c_n\}$, $n=1, 2, \ldots$ На каждом i-м шаге к одному кластеру относятся те объекты, расстояние между которыми не превосходит c_i.

При выполнении процедуры кластеризации происходит построение иерархического классификационного дерева. Под ним понимается множество разбиений исходной выборки на кластеры, упорядоченное по уровням иерархии или, что то же самое, по номеру шага иерархической процедуры.

Следует отметить, что привлекательность иерархических процедур состоит в возможности полного и достаточно тонкого анализа структуры исследуемого множества объектов. Действительно, имея в распоряжении иерархическое дерево, можно решать самые разнообразные задачи классификации. Например, определить заранее неизвестное число классов или, наоборот, выполнить разбиение на заданное число классов, провести анализ однородности полученных кластеров, выбрать оптимальную классификацию (с учетом поставленной задачи). Удобным свойством подобных

алгоритмов является возможность наглядной (визуальной) интерпретации проведенного анализа.

В настоящей работе предлагается «интеллектуальная» модель реализации кластерного анализа, включающая в себя модуль предварительного анализа исходных данных с целью определения их типа (номинальный, бинарный, количественный). Далее, в зависимости от установленного типа данных, реализуется этап выбора способа определения расстояния между двумя объектами (в терминологии статистических пакетов — реализациями).

Метрика выбирается из числа наиболее подходящих для задач кластерного анализа в зависимости от типа исходных данных [2]:

Количественный тип:

Евклидово расстояние, квадратичное Евклидово расстояние, косинус угла между двумя векторами значений, расстояние Чебышева.

Номинальный тип:

χ^2 (Хи-квадрат), φ^2 (Фи-квадрат), расстояние Хэмминга.

Кроме того, для номинальных шкал, как очень естественный, предлагается следующий подход: мерой отклонения между двумя реализациями можно считать отношение числа несовпадающих значений переменных этих реализаций к числу положительных переменных, кроме пропущенных.

Бинарный тип:

Евклидово расстояние, квадратичное Евклидово расстояние, расстояние Хэмминга, расстояние Жаккара, расстояние Хаманна.

При конструировании кластер-процедур возникает также понятие расстояния не только между отдельными объектами, но между группами объектов. В работе используются следующие подходы к определению расстояния, характеризующего взаимное расположение отдельных групп объектов (область применения каждой из функций межклассового расстояния определяется различными типами исходных данных и различных кластер-процедур):

- ближний сосед — расстояние между двумя группами равно расстоянию между ближайшими объектами из этих групп;
- центры тяжести — расстояние между двумя группами равно расстоянию между их математическими ожиданиями;
- дальний сосед — расстояние между двумя группами равно расстоянию между самыми дальними объектами из этих групп;
- среднее — расстояние между двумя группами равно среднему арифметическому всевозможных попарных расстояний между представителями рассматриваемых групп.

Приступая к выбору расстояния между группами, необходимо учесть следующий немаловажный факт: каждый конкретный вид иерархической процедуры отыскивает в исходной выборке не те группы, которые там реально существуют, а те, для поиска которых он предназначен. Это проявляется как следствие того, что круг задач, при решении которых не происходит принудительного навязывания данным заранее предопределенной структуры, достаточно узок. В [3, 4] приводятся тестовые примеры применения к выборке различных алгоритмов.

Расчеты показывают, что использование алгоритма ближнего соседа приводит к разбиению на классы, в корне отличному от результата при-

менения алгоритма дальнего соседа. При этом, фактически, структура данных определяет выбор алгоритма. Причиной непригодности алгоритма ближнего соседа для конкретной выборки может оказаться наличие цепочки близких друг другу объектов, соединяющих первую и вторую группы. Нетрудно видеть, что это связано с общим свойством алгоритма ближнего соседа. Получаемые в результате объединения (разбиения) кластеры могут иметь произвольную (необязательно выпуклую) форму. Тем не менее, это, конечно, не может объявляться недостатком алгоритма. Просто его следует применять в тех случаях, когда эта особенность не будет приводить к неестественной классификации.

В случае сложных форм естественных группировок как раз цепочный эффект алгоритма ближнего соседа обеспечивает верную классификацию. В то же время алгоритм дальнего соседа ориентирован на поиск скоплений типа шаровых. При этом, если исходные данные образуют естественные группы сложной формы, следует ожидать навязывания данным шаровой структуры, предполагаемой алгоритмом, а значит, и неверной классификации.

Промежуточными, т. е., ориентированными на поиск групп данных не очень сложной, но и не шаровой формы, являются алгоритмы центров тяжести и средней связи. Они, в основном, предназначены для данных, имеющих форму типа гиперэллипсоидной. Следует отметить, что алгоритм центров тяжести ближе к алгоритму дальнего соседа, а алгоритм средней связи может находить, как и процедура ближнего соседа, группировки не выпуклой формы.

Иерархические процедуры типа средней связи или ближнего соседа могут успешно применяться для выявления естественных групп признаков. Обнаружение таких групп позволяет снизить размерность пространства признаков путем выбрасывания близких в смысле введенной меры близости признаков. Каждую группу признаков можно заменять новым признаком, обладающим общим для этой группы свойством и соответствующим реальной интерпретации обрабатываемых данных. В этом случае именно иерархическая процедура будет работать успешно.

Предлагаемая программная реализация позволяет отслеживать процесс построения системы кластеров на всех этапах на основе непосредственного изучения дендрограммы, останавливая его в случае достижения необходимого результата. Если же структура иерархического дерева вызывает сомнения, возможно варьирование способов выбора как метрики, определяющей расстояние между исходными объектами, так и способа определения расстояния между кластерами. Одним из возможных способов проверки устойчивости результатов кластерного анализа может быть метод сравнения результатов, полученных для различных алгоритмов кластеризации.

Для иллюстрации устойчивой структуры иерархического дерева (на определенном уровне) предлагается результат кластерного анализа данных по выборам Президента РФ в 1996 г. по южным районам Тюменской области.

В качестве основной цели классификации была выбрана последовательная группировка вплоть до получения четырех кластеров.

Исходные данные:

Район	% принявших участие	Зюганов	Ельцин	Жириновский	Лебедь	Явлинский	Против всех
Абатский	78,32	34,50	32,19	11,44	11,38	3,58	2,19
Армизонский	82,51	40,64	29,56	13,95	7,82	3,08	1,78
Аромашевский	75,06	36,82	30,30	15,16	9,85	3,00	1,50
Бердюжский	79,72	42,40	29,34	9,95	9,00	3,08	1,75
Вагайский	77,39	44,27	32,30	8,37	6,79	3,24	1,52
Викуловский	76,85	38,78	24,64	12,45	15,10	3,54	1,88
Голышмановский	70,32	37,13	28,95	17,15	8,98	3,03	1,72
Исетский	72,33	31,80	39,65	10,19	10,68	2,50	1,85
Казанский	75,57	35,26	32,30	13,05	10,05	3,68	1,92
Нижне-тавдинский	75,81	33,87	36,41	10,46	11,05	3,14	1,48
Омутинский	71,65	36,05	27,62	19,41	9,62	3,33	1,24
Сладковский	82,66	43,74	23,36	14,51	9,75	3,59	1,94
Сорокинский	76,56	36,2	27,31	17,09	8,55	4,37	2,10
Тобольский	76,31	40,73	32,59	7,56	9,37	3,72	2,00
Тюменский	67,23	25,69	39,72	8,85	16,08	4,36	1,75
Уватский	73,86	17,78	51,02	10,04	10,03	5,16	2,14
Упоровский	74,07	32,85	30,94	16,29	10,33	3,78	2,15
Юргинский	79,21	39,01	28,44	13,77	10,59	2,74	2,19
Ярковский	74,71	39,69	35,26	8,77	7,59	3,02	1,66

Результат с метрикой Евклидово расстояние, расстояние между группами — ближний сосед:

Номер кластера	Район
1.	Уватский
2.	Тюменский
3.	Сладковский
4.	Все остальные

Результат с метрикой Евклидово расстояние, расстояние между группами — дальний сосед:

Номер кластера	Район
1.	Уватский
2.	Тюменский, Исетский
3.	Сладковский, Викуловский, Юргинский, Армизонский
4.	Все остальные

Результат с метрикой Евклидово расстояние, расстояние между группами — средняя связь:

Номер кластера	Район
1.	Уватский
2.	Тюменский
3.	Сладковский, Викуловский, Юргинский, Армизонский, Тобольский, Вагайский, Ярковский, Бердюжский
4.	Все остальные

Предлагаемый метод будет предпочтителен также там, где имеется очень много входных данных, в которых скрыты закономерности. В этом случае можно учесть различные взаимодействия между показателями-признаками, характеризующими такие данные. Это особенно важно в системах обработки информации (распределенных базах данных, телекоммуникационных и экспертных системах), в частности, для ее предварительного анализа или отбора, выявления «выпадающих» фактов или грубых ошибок человека, принимающего решения. Целесообразно использовать предлагаемый метод в задачах с неполной информацией, а также в тех случаях, когда решение можно подтвердить интуитивными соображениями.

ЛИТЕРАТУРА

1. Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере. М.: ИНФРА-М., 1998. 528 с.

2. Справочник по прикладной статистике / Под ред. Э. Ллойда, У. Ледермана, Ю. Н. Тюрина. М.: Финансы и статистика, 1990. 830 с.

3. Александров В. В., Горский Н. Д. Алгоритмы структурного метода обработки данных. Л.: Наука, 1983. 208 с.

4. Александров В. В., Алексеев А. И., Горский Н. Д. Анализ данных на ЭВМ. (На примере системы СИТО). М.: Финансы и статистика, 1990. 192 с.

Галина Викторовна РУБЛЕВА — старший преподаватель кафедры математического анализа и теории функций математического факультета

УДК 519.2 (075.8)

ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ ПАКЕТА «АНАЛИЗ ДАННЫХ» (MICROSOFT EXCEL)

АННОТАЦИЯ. В статье предложен вариант дополнения к «Справке» пакета «Анализ данных» (Microsoft Excel) по дисперсионному анализу.

In the article the supplement to «Information» of the package «The Analysis of the data» (Microsoft Excel) on dispersing analysis is offered.

В настоящее время существуют пакеты прикладных программ (Microsoft Excel — пакет «Анализ данных», статистические функции, финансовые функции, Statistica, SPSS), позволяющие выполнять расчеты для изучения взаимосвязей между экономическими явлениями и процессами, построения прогнозов в экономике и социологии.

Соответствующая литература представляет собой справочную информацию, предназначенную для специалистов, и не может быть использована в качестве учебных пособий по таким дисциплинам, как «математическая статистика», «эконометрика», «прикладная статистика». Поэтому представляется целесообразным издание специальной методической литературы, содержащей необходимые теоретические сведения, четкие указания по вводу нужной информации и пояснения к выходным данным.