УДК 631.474:631.482.1 (571.16)

ИСПОЛЬЗОВАНИЕ ГАРАНТИЙНОГО МОМЕНТА ОСТАНОВКИ В ЗАДАЧАХ КЛАСТЕРИЗАЦИИ ЭКОЛОГИЧЕСКИХ ДАННЫХ

© Т.Э. Королева, А.Г. Закиров, Л.Л. Фролова, Б.Р. Григорьян

Koroleva T.E., Zakirov A.G., Frolova L.L., Grigoryan B.R. Using the moment of stopping in the tasks of cluster analysis for ecological data. The article contains a comparison between the existing statistical methods and the one developed by the authors.

Задача прогнозирования динамики состояния экологических объектов без проведения дорогостоящих исследований в последнее время становится все более актуальной. В этих целях может быть использована группировка с помощью универсальных методов кластерного анализа. Для многих исследователей стало очевидным, что экологические данные необходимо обрабатывать, имея в виду нарушения предположения нормальности для всего комплекса данных. В этом случае выбор аппарата кластерного анализа и непараметрических методов [4, 7] представляется вполне оправданным в задачах группировки реальных данных. Однако для прогнозных выводов очень важно решение вопроса о репрезентативности данных. Однозначный ответ на этот вопрос можно получить только используя параметрические методы. Для этих методов развит мощный математический аппарат, который, при правомерном применении, дает достаточно устойчивые результаты.

В задачах кластеризации очень важным является вопрос о том, что считать критерием достаточного разбиения. Очевидным критерием качества и обоснованности полученного разбиения является содержательный анализ результатов, основанный на осмыслении исследователем возможных причинных механизмов обособления полученных групп объектов. Статистические критерии оказываются вспомогательными инструментами в процессе такого анализа [1]. Тем не менее, формальное решение этого вопроса представляет большой практический интерес. В качестве критерия остановки разбиения авторами предлагается использовать процедуру подсчета необходимого объема выборки [8, 9], основанную на определении гарантийного момента остановки.

В результате кластеризации число данных в получаемых группах, как правило, небольшое, но данные достаточно однородны. Для правомерности применения параметрических методов всегда необходимо определить, можно ли приблизить эмпирическое распределение теоретическим [10]. Так, для получения достоверных выводов проверяется согласованность данных кластера с законом нормального распределения. В статистических пакетах используются стандартные процедуры: проверка нормальности с помощью асимметрии и эксцесса, критерия χ^2, а также критериями типа Колмогорова – Смирнова, статистики которых обладают более быстрой сходимостью к предельным распределениям, т. е. их можно применять для малых выборок. Считается, что мощность критериев типа Колмогорова – Смирнова, вообще говоря, выше, чем у критерия χ^2 [3].

Проблема получения устойчивых результатов при обработке выборок небольшого объема [5, 6] была и остается актуальной для исследователей, работающих в области биологии, медицины и некоторых других естественных наук. Стандартные критерии и методы параметрической статистики предполагают, что объем выборки должен быть ≥ 25 наблюдений (и тогда выборка называется сравнительно небольшой). Но, как известно исследователям-биологам, собрать даже такой массив данных за короткое время летних экспедиций трудно, и при этом требуются большие затраты. В таких случаях исследователи поступают двояко: используют непараметрические тесты или, пренебрегая предположениями о нормальности, применяют параметрические методы. Отметим, что использование непараметрических методов часто неудобно, так как обычно эти критерии имеют меньшую мощность, чем параметрические, и обладают меньшей гибкостью. В прикладных расчетах использование стандартных параметрических и непараметрических методов определения качества разбиения выборки обладает одним существенным недостатком, – они не дают конкретного значения объема выборки данных, обеспечивающего репрезентативность статистик и устойчивость разбиения.

Очень удобно на практике использовать процедуру определения минимального объема выборки, основанную на вычислении гарантийного момента остановки [8, 9], которая указывает объем данных, необходимых для достоверных статистических выводов, по имеющимся данным. Процедура дает возможность определить, является ли группа данных репрезентативной или нет, и могут ли они быть описаны параметрически, так как принадлежность к закону нормального распределения является необходимым условием применения этой процедуры. Метод гарантийного момента остановки реализован в виде программной процедуры «Объем выборки», используемой в пакете «Сервис Base». Для проверки практической применимости этой процедуры были проведены сравнительные расчеты по проверке малых выборок на нормальность с использованием стандартных тестов Шапиро – Уилкса и Колмогорова – Смир-

Таблица 1

Тип и подтип почвы	Металл	Объем выборки	Процедура «Объем выборки»						Проверка на нормальность тестами	
			5 %	10 %	15 %	20 %	25 %	30 %	Шапиро – Уилкса	Колмогорова – Смирнова
Корич. серая лесная	Pbv	13	106	27	12	7	5	3	0,0611545 (+)	0,383143 (+)
	Pbp	13	322	81	36	21	13	9	0,504126 (+)	0,948208 (+)
Темно-серая лесная	Pbv	8	96	24	11	6	4	3	0,656056 (+)	0,999997 (+)
	Pbp	10	401	101	45	26	17	12	0,100488 (+)	0,836255 (+)
Черноземн. Тип	Pbv	8	44	11	5	3	2	2	0,60509 (+)	0,994132 (+)
	Pbp	9	234	59	26	15	10	7	0,0820732 (+)	0,547587 (+)
Дерновая	Pbv	6	104	26	12	7	5	3	0,0283889(-)	0,727381 (+)
	Pbp	7	224	56	25	14	9	7	0,680225 (+)	0,971808 (+)
Черн. оподз.	Pbv	19	59	15	7	4	3	2	0,272127 (+)	0,743984 (+)
	Pbp	21	757	190	85	48	31	22	0,0000057 (-)	0,181074 (+)
Черн. выщ.	Pbv	22	45	12	5	3	2	2	0,634036 (+)	0,97063 (+)
	Pbp	22	327	82	37	21	14	10	0,0062516 (-)	0,543918 (+)
Св. серая лесная	Pbv	32	99	25	11	7	4	3	0,0535257 (+)	0,477602 (+)
	Pbp	33	529	133	59	34	22	15	2,0306E-7 (-)	0,0061222 (-)
Серая лесная	Pbv	33	65	17	8	5	3	2	0,0005266 (+)	0,45837 (+)
	Pbp	37	559	140	63	35	23	16	0,0008975 (-)	0,140217 (+)

Таблица 2

Сравнение результатов теста Шапиро – Уилкса и процедуры «Объем выборки»

Тип и подтип почвы	Металл	10 %	15 %	20 %	25 %	30 %
Корич. серая	Pbv	-	+	+	+	+
	Pbp	-	-	-	+	+
Темн. серая	Pbv	-	-	+	+	+
	Pbp	-	-	-	-	-
Черн. тип	Pbv	-	+	+	+	+
	Pbp	-	-	-	-	+
Дерновая	Pbv	+	+	+	-	-
	Pbp	-	-	-	-	+
Черн. оподз.	Pbv	+	+	+	+	+
	Pbp	+	+	+	+	+
Черн. выщ.	Pbv	+	+	+	+	+
	Pbp	+	+	-	-	-
Св. серая	Pbv	+	+	+	+	+
	Pbp	-	+	+	-	-
Серая	Pbv	+	+	+	+	+
	Pbp	+	+	-	-	-
Совпадений	число	8	11	10	9	11
	Процент	50 %	68,8 %	62,5 %	46,3 %	68,8 %

Таблица 3

Сравнение результатов теста Колмогорова – Смирнова и процедуры «Объем выборки»

Тип и подтип почвы	Металл	10 %	15 %	20 %	25 %	30 %
Корич. серая	Pbv	+	+	+	+	+
	Pbp	-	-	-	+	+
Темн. серая	Pbv	-	-	+	+	+
	Pbp	-	-	-	-	-
Черн. тип	Pbv	-	+	+	+	+
	Pbp	-	-	-	-	+
Дерновая	Pbv	-	-	-	+	+
	Pbp	-	-	-	-	+
Черн. оподз.	Pbv	+	+	+	+	+
	Pbp	-	-	-	-	-
Черн. выщ.	Pbv	+	+	+	+	+
	Pbp	-	-	+	+	+
Св. серая	Pbv	+	+	+	+	+
	Pbp	-	-	-	+	-
Серая	Pbv	+	+	+	+	+
	Pbp	-	-	+	+	+
Совпадений	число	5	6	9	12	13
	Процент	31,3 %	37,5 %	46,3 %	75 %	81 %

нова, реализованных в пакете «Statgraphics Plus». Для проверки использовались данные по концентрации валовой и подвижной форм свинца в различных типах почв Предволжья (данные предоставлены Институтом экологии природных систем Академии наук Республики Татарстана). Результат работы процедуры «Объем выборки» представлен оценкой размера устойчивого объема данных при заданной полуширине (5–30 % от среднего) доверительного интервала для среднего. В таблице 1 показаны результаты работы процедуры «Объем выборки» и проверки на принадлежность нормальному закону распределения двумя статистическими тестами данных по содержанию валового и подвижного свинца в различных типах почв.

Для процедуры «Объем выборки» гипотеза о нормальности исходных данных принимается, если вычисленный репрезентативный размер выборки меньше или равен объему выборки исходных данных. Как видно из таблицы 1, процедура «Объем выборки» в большинстве случаев согласуется с проверкой данных на принадлежность к закону нормального распределения при ширине доверительного интервала 20–30 % от среднего. Кроме того, она указывает необходимый для устойчивых выводов объем данных. Ширина доверительного интервала 20–30 % от среднего свидетельствует о низкой точности данных, что характерно практически для всех экологических данных. Процедура чувствительна к точности данных, что видно при изменении ширины доверительного интервала. Степень совпадения выводов о нормальности данных, полученных процедурой «Объем выборки» и тестами Шапиро – Уилкса и Колмогорова – Смирнова, показана в таблицах 2 и 3 соответственно.

Таблица 4

Сравнение результатов теста Шапира – Уилкса и теста Колмогорова – Смирнова

Тип и подтип почвы	Металл	Совпадение выводов тестов
Корич. серая лесная	Pbv	+
	Pbp	+
Темн. серая лесная	Pbv	+
	Pbp	+
Черн. тип	Pbv	+
	Pbp	+
Дерновая	Pbv	–
	Pbp	+
Черн. оподз.	Pbv	+
	Pbp	–
Черн. выщ.	Pbv	+
	Pbp	-
Св. серая лесная	Pbv	+
	Pbp	+
Серая лесная	Pbv	+
	Pbp	–
Совпадений	число	12
	процент	75 %

Знак «+» означает совпадение выводов соответствующего критерия и процедуры «Объем выборки», например, если гипотеза о нормальности отвергается и процедура признает объем выборки нерепрезентативным. Рассогласование в выводах обозначено знаком «–»

Из сравнительного анализа таблиц видно, что при полуширине доверительного интервала 30 % от среднего степень совпадения выводов процедуры «Объем выборки» с выводами критерия Колмогорова – Смирнова равно 81 %, а с критерием Шапиро – Уилкса – 69 %. Совпадение между самими критериями составляет 75 %.

Результаты расчетов и их анализ показывают, что процедуру определения гарантийного момента остановки можно использовать при практических построениях группировок экологических данных. Кроме того, эта процедура дает полезную предварительную информацию о возможности параметрического описания имеющегося набора данных и объеме выборки, необходимом для получения устойчивых выводов при дальнейшей статистической обработке.

ЛИТЕРАТУРА

1. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983. 471 с.
2. *Закс Ш.* Теория статистических выводов. М.: Мир, 1975. 776 с.
3. Справочник по прикладной статистике / Под ред. Э. Ллойда, У. Ледермана. Т. 1 и 2. М.: Финансы и статистика, 1989.
4. *Райзин Дж.Вэн* Классификация и кластер. М.: Мир, 1980. 389 с.
5. *Петров А.А.* Проверка статистических гипотез о типе распределения по малым выборкам // Теория вероятностей и ее применения. 1956. Т. 1. Вып. 2. С. 248-269.
6. *Володин И.Н.* Проверка гипотезы нормальности распределения по малым выборкам (многомерный случай). Казань: Изд-во Казан. ун-та, 1964. С. 21-25.
7. *Тюрин Ю.Н.* Непараметрические методы статистики. М.: Знание, 1978. 64 с.
8. *Закиров А.Г., Королева Т.Э., Фролова Л.Л.* К оценке репрезентативности экологических данных // Казанский мед. ж. 1992. Т. 73. № 4. С. 295-298.
9. *Frolova L.L., Zakirov A.G., Koroleva T.E.* The evaluation of data representation / The second UK Congress of Biotechnology (Biotechnology'94). Proceedings of Second Conference on Advances in Biochemical Engineering, Brighton, UK, 4-6 July 1994. P. 160-162.
10. *Орлов А.И.* О критериях согласия с параметрическим семейством // Заводская лаборатория. Т. 61. № 7. 1995. С. 59-61.

Поступила в редакцию 20 сентября 2000 г.