ОЦЕНКА УСТОЙЧИВОСТИ ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ НА ОСНОВЕ КОФЕНЕТИЧЕСКОЙ КОРРЕЛЯЦИИ

А.Ю. Тимофеева, С.Б. Цыренжапова
Новосибирский государственный технический университет,
Россия, г. Новосибирск, пр. К.Маркса, 20, 630073
E-mail: a.timofeeva@corp.nstu.ru, tsyrenzhapova.sb@gmail.com

ASSESSMENT OF THE ROBUSTNESS OF HIERARCHICAL CLUSTERING BASED ON COPHENETIC CORRELATION

A.Yu. Timofeeva, S.B. Tsyrenzhapova
Novosibirsk State Technical University, Russia, Novosibirsk, Prospekt K. Marksa, 20, 630073
E-mail: a.timofeeva@corp.nstu.ru, tsyrenzhapova.sb@gmail.com

Abstract. *If there are any outliers in the data, hierarchical clustering can produce poor results. In addition, the dendrogram is sensitive to a set of characteristics by which objects are compared. Therefore, it is required to investigate two types of robustness of hierarchical clustering – to a set of objects and to a set of characteristics. For this, an original approach based on the use of bootstrapping is proposed. As an internal criterion for the reliability of hierarchical clustering, the cophenetic correlation coefficient is used. In the simulation study, various methods of hierarchical clustering are compared for robustness of two types. Recommendations are given on the applicability of methods of hierarchical clustering.*

Введение. При решении задач разбиения объектов, описываемых множеством признаков, на однородные группы для визуального представления результатов часто используется дендрограмма. Она представляет собой дерево, отражающее близость исследуемых объектов, которая определяется на основе различных мер сходства (различия). Для построения такого дерева применим целый набор методов. На практике аналитику необходимо выбрать меру близости и метод построения дендрограмм. Так в работе [1] в качестве критерия выбора предлагается кофенетический коэффициент корреляции, выступающий мерой близости попарных расстояний между объектами. При расчете этого коэффициента исходная матрица расстояний сопоставляется с воспроизведенной с помощью дендрограммы. Авторами статьи [1] смоделированы две ситуации (без выбросов и с выбросами) и рекомендованы лучшие методы кластеризации для разных мер расстояний. Ими оказались метод средней связи и центроидный метод, обеспечивающие высокие значения кофенетической корреляции в обеих ситуациях.

Однако используемый критерий качества построения дендрограммы не гарантирует устойчивость полученных результатов. Иными словами, с его помощью не удается обнаружить, что в выборочных данных есть искажающие наблюдения и идентифицировать их. Наоборот, при наличии выбросов коэффициент кофенетической корреляции закономерно больше, чем для аналогичной модели без аномальных наблюдений. Эта особенность отмечена в работах [1, 2] и связана с тем, что дендрограмма подстраивается под резко выделяющиеся наблюдения, сильно удаленные от остальных точек. Значит, при высоких значениях кофенетической корреляции есть опасность ложного вывода об отличном качестве дендрограммы, обусловленного наличием аномальных наблюдений.

Подходы к исследованию устойчивости. В этой связи предлагается другой критерий качества, характеризующий дендрограмму с точки зрения устойчивости результатов. Он так же основан на кофенетической корреляции, но сопоставление производится на псевдовыборках, сформированных с использованием бутстрэппинга. Самый простой вариант построения псевдовыборок – это отбрасывание одного наблюдения (признака). В случае если исключается резко выделяющееся наблюдение (или признак), очевидно, построенная дендрограмма будет сильно отличаться от исходной, что свидетельствует о неустойчивости результатов. Тем самым для исследования устойчивости к набору признаков разработан следующий подход.

Шаг 0. По исходным данным строится дендрограмма.

Шаг 1. Из набора данных исключается j-й признак. Строится дендрограмма по новым данным.

Шаг 2. Вычисляется кофенетический коэффициент корреляции между дендрограммами, построенными на шагах 0 и 1.

Шаги 1 и 2 повторяются для всех $j = 1, \ldots, m$, где m – число признаков. Тем самым получаем набор из m коэффициентов кофенетической корреляции. Если эти коэффициенты практически не меняются, то можно дать заключение об устойчивости результатов, в противном случае, наиболее выделяющиеся коэффициенты указывают на признаки, которые можно считать аномальными.

Некоторую сложность представляет расширение этого подхода для исследования устойчивости к множеству объектов. Дело в том, что дендрограммы сопоставимы только при одинаковом числе объектов. Значит, нельзя вычислить кофенетическую корреляцию между дендрограммами, построенными по исходным данным и с удалением одного наблюдения. Поэтому для сопоставимости предлагается удалять из исходной выборки ближайшего соседа. Ближайший сосед определяется на основе той же меры близости, что и при построении дендрограммы. Тогда подход к исследованию устойчивости к множеству объектов можно представить следующим образом.

Шаг 0. Вычисляется матрица расстояний между объектами, для каждого объекта определяется ближайший сосед.

Шаг 1. Из исходного множества объектов удаляется ближайший к i-му объекту сосед. По полученным данным строится дендрограмма.

Шаг 2. В исходных данных i-й объект заменяется на его ближайшего соседа, ближайший сосед удаляется из данных. Строится дендрограмма по новому набору данных.

Шаг 3. Вычисляется кофенетический коэффициент корреляции между дендрограммами, построенными на шагах 1 и 2.

Шаги 1-3 повторяются для всех $i = 1, \ldots, n$, где n – число объектов.

Результаты экспериментов. Для сравнения различных методов иерархического кластерного анализа на основе предложенных подходов проведены вычислительные эксперименты. Для исследования устойчивости к множеству объектов наблюдения моделировались с использованием нормальной смеси с функцией распределения вида $F(x) = (1-\lambda)\Phi(x;0,\sigma^2) + \lambda\Phi(x;0,k\sigma^2)$, где $\Phi(x;0,\sigma^2)$ – функция нормального распределения с нулевым математическим ожиданием и дисперсией σ^2. Величина $k > 1$ определяет степень неоднородности данных, связанную с наличием выбросов;

$k = 10$, $\sigma^2 = 1$. Параметр $\lambda \in [0,1]$ характеризует степень засорения данных аномальными наблюдениями и задан равным 0,05.

Для исследования устойчивости к набору признаков моделировались совокупности из десяти признаков. Значения первого признака X_1 генерировались из стандартного нормального распределения. Значения остальных, кроме последнего, определялись из соотношения $X_j = X_1 + 0.5\varepsilon_j$, $j = 2,...,9$, где ε_j – независимые случайные величины, имеющие стандартное нормальное распределение. Тем самым девять признаков коррелировали друг с другом. Последний признак представлял собой случайный шум.

Число объектов во всех экспериментах взято равным 100. Результаты усреднялись по 500 повторениям. В качестве меры различия использовалось Евклидово расстояние. В таблице 1 представлена сводная информация по кофенетическим коэффициентам корреляции, рассчитанным с помощью пакета dendextend статистической среды R [3].

Таблица 1

Результаты исследования устойчивости методов иерархической кластеризации

Методы	Устойчивость к множеству объектов		Устойчивость к набору признаков			
	Среднее	Минимум	Медиана	Минимум	Квартиль 25%	Квартиль 75%
Средней связи	0,996	0,269	0,802	0,379	0,664	0,903
Центроидный	0,997	0,201	0,849	0,196	0,737	0,932
Полной связи	0,971	0,193	0,549	0,174	0,409	0,773
Медианный	0,969	0,040	0,396	-0,043	0,221	0,608
Одиночной связи	0,997	0,085	0,942	0,357	0,913	0,962
Уорда	0,984	0,399	0,748	0,286	0,538	0,878

Заключение. При исследовании устойчивости к множеству объектов кофенетическая корреляция в большинстве случаев близка к единице, поэтому судить о неустойчивости результатов следует по минимальным значениям. Наибольшее из них обеспечивается при использовании метода Уорда. При изменении набора признаков коэффициент кофенетической корреляции больше варьируется, поэтому его изменчивость можно охарактеризовать, например, с помощью межквартильного размаха. Наименьшее его значение достигается при применении метода одиночной связи. Таким образом, методы Уорда и одиночной связи могут быть рекомендованы как наиболее устойчивые методы иерархического кластерного анализа.

Работа поддержана грантом Министерства образования и науки РФ в рамках проектной части государственного задания, проект № 2.2327.2017/4.6 «Интеграция моделей представления знаний на основе интеллектуального анализа больших данных для поддержки принятия решений в области программной инженерии».

СПИСОК ЛИТЕРАТУРЫ

1. Saraçli S., Doğan N., Doğan İ. Comparison of hierarchical cluster analysis methods by cophenetic correlation // Journal of Inequalities and Applications. – 2013. – 2013:203.
2. Johnson R.A., Wichern D.W. Applied Multivariate Statistical Analysis. – New York: Prentice Hall, 2002.
3. Galili T. Dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering // Bioinformatics. – 2015. – Т. 31. – № 22. – С. 3718-3720.