Математические методы исследования

УДК 519.234

УСТОЙЧИВОСТЬ КЛАССИФИКАЦИИ ОТНОСИТЕЛЬНО ВЫБОРА МЕТОДА КЛАСТЕР-АНАЛИЗА

© А. И. Орлов[1]

Статья поступила 13 сентября 2011 г.

Впервые на примере продемонстрирована устойчивость результатов классификации относительно выбора метода кластер-анализа. Одни и те же данные подробно проанализированы методами «ближнего соседа» и «дальнего соседа». Показано, что полученные результаты близки.

Ключевые слова: теория классификации; метод «ближнего соседа»; метод «дальнего соседа»; устойчивость.

В теории устойчивых математических методов и моделей [1, 2] есть следующее общее утверждение: «Можно рекомендовать обрабатывать данные несколькими способами (методами). Выводы, общие для всех способов, скорее всего, отражают реальность (являются объективными). Выводы, меняющиеся от метода к методу, субъективны, зависят от исследователя, выбравшего тот или иной метод анализа данных. Здесь речь идет об устойчивости выводов по отношению к выбору метода» [3].

Применим это общее утверждение в теории классификации [4 – 7], насколько нам известно, впервые. Для построения классификаций часто применяется так называемый агломеративный иерархический алгоритм «Дендрограмма», в котором вначале все элементы рассматриваются как отдельные кластеры, а затем на каждом шагу объединяются два наиболее близких кластера.

Для работы «Дендрограммы» необходимо задать правило вычисления расстояния между кластерами. Оно вычисляется через показатель различия $d(x, y)$ классифицируемых объектов x и y. В качестве примера в табл. 1 приведены показатели различия для совокупности из 17 объектов. Поскольку $d(x, x) = 0$ для любого x и $d(x, y) = d(y, x)$ для любых x и y, то квадратная матрица, задающая попарные показатели различия, является симметричной относительно главной диагонали, а на главной диагонали стоят нули. Поэтому достаточно привести лишь часть матрицы, находя-

[1] Кафедра теории классификации Международного университета междисциплинарных знаний, Москва, Россия; e-mail: prof-orlov@mail.ru

Таблица 1. Показатели различия между объектами

Номера объектов	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	7	22	19	20	51	23	27	25	43	52	63	46	47	46	51	57
2		14	50	30	50	30	29	26	42	61	64	47	53	57	61	60
3			46	41	53	35	40	38	38	61	77	70	59	66	55	65
4				30	57	35	37	37	50	45	54	52	54	49	64	56
5					18	19	31	31	37	43	52	40	64	52	54	61
6						41	60	60	55	61	71	80	61	80	63	84
7							25	41	47	60	59	57	68	70	53	55
8								40	41	47	60	59	57	68	70	53
9									66	59	53	56	50	50	62	57
10										35	52	35	36	52	50	32
11											14	43	62	70	65	51
12												39	55	63	60	71
13													67	72	51	46
14														24	38	66
15															8	38
16																54

щуюся выше главной диагонали, как и показано в табл. 1.

В прикладных исследованиях показатели различия определяют по статистическим данным или находят с помощью экспертного исследования. Если классифицируемые объекты описываются конечномерными векторами одной и той же размерности, то в качестве показателя различия можно применять то или иное расстояние в конечномерном пространстве — евклидово, блочное и др. Если классифицируемые объекты — признаки, то показатель различия можно получить, отнимая от единицы тот или иной коэффициент парной корреляции — линейный коэффициент Пирсона, непараметрический коэффициент ранговой корреляции Спирмена или Кендалла. Если классифицируемые объекты — бинарные отношения, то в качестве показателя различия можно использовать расстояние Кемени или расстояние подобия. В общем случае нет основания считать, что выполнено неравенство треугольника; этим показатель различия отличается от метрики (расстояния). Иногда вместо термина «показатель различия» используется термин «мера близости». Последний представляется неудачным, поскольку чем ближе друг к другу объекты, т.е. чем «близость» больше, тем «мера близости» меньше.

Различные варианты агломеративного иерархического алгоритма «Дендрограмма» различаются правилами вычисления расстояния между кластерами.

В алгоритме «ближнего соседа» расстоянием между кластерами называется минимальный из показателей различия между парами объектов, один из которых входит в первый кластер, а другой — во второй. В алгоритме «дальнего соседа» — это максимальный из показателей различия между парами объектов, один из которых входит в первый кластер, а другой — во второй. В алгоритме средней связи данное расстояние рассчитывается как средняя связь (отсюда и название), т.е. как среднее арифметическое показателей различия между парами объектов, один из которых входит в первый кластер, а другой — во второй.

Для этих трех алгоритмов после ряда шагов все объекты объединяются вместе; результат работы алгоритма представляет собой дерево (в терминах теории графов) последовательных объединений, так называемую «Дендрограмму». Из нее можно выделить кластеры разными способами: исходя из заданного числа кластеров; из соображений предметной области; исходя из устойчивости (если разбиение долго не менялось при возрастании порога объединения, то оно отражает реальность) и т.д.

Кроме трех описанных алгоритмов, наиболее часто используемых при решении практических задач, имеется бесконечно много иных вариантов агломеративного иерархического алгоритма «Дендрограмма». Во-первых, если $d(x, y)$ — показатель различия, то его степень $d^a(x, y)$ при любом $a > 0$ также является показателем различия, поэтому алгоритмов рассматриваемого вида столько же, сколько точек на прямой. Во-вторых, практически в любом конкретном пространстве существует весьма много показателей различия различных видов [8].

Если классы реальны [7], естественны, существуют на самом деле (а не только в сознании исследователя), четко отделены друг от друга, то любой алгоритм кластер-анализа их выделит. Следовательно, *в качестве критерия естественности классификации следует рассматривать устойчивость относительно выбора алгоритма кластер-анализа*.

Проверить устойчивость можно, применив к данным несколько подходов, например, такие столь непохожие алгоритмы, как «ближнего соседа» и «дальнего соседа» (для множества показателей различия они соответствуют $a = -\infty$ и $a = +\infty$). Если полученные результаты содержательно близки, то они адекватны действительности. В противном случае следует предположить, что естественной классификации не существует, задача кластер-анализа не имеет решения и можно проводить только группировку [7]. Чтобы продемонстрировать возникающие при этом эффекты, применим алгоритмы «ближнего соседа» и «дальнего соседа» к данным табл. 1.

При практическом построении дерева необходимо выделить расстояния (показатели различия) между кластерами, при которых происходит объединение тех или иных кластеров. Для этого достаточно (как при ручном, так и при машинном счете) упорядочить содержащиеся в табл. 1 числа в порядке возрастания и, просматривая их, отмечать, происходит при каждом из них объединение кластеров или нет. Полученные разбиения на кластеры (классификации) приведены в табл. 2. Объекты (обозначены номерами от 1 до 17), входящие в один кластер, выделены фигурными скобками; для упрощения записей кластеры из одного элемента скобками не выделяются.

Сравнивая дендрограммы в целом, а не отдельные получаемые из них классификации (разбиения на кластеры), можно сделать следующие выводы.

1. Двум методам соответствует разное число объединений — 12 для метода «ближнего соседа» и 15 для метода «дальнего соседа». Это связано с тем, что в методе «ближнего соседа» при одном и том же расстоянии между кластерами происходит не одно объединение, а больше. Например, при расстоянии 14 образовались два объединения — элемент 3 присоединился к кластеру {1, 2} и одновременно элементы 15 и 16 объединились в кластер.

2. Дерево для метода «дальнего соседа» более чем в два раза «выше», чем дерево для «ближнего соседа»: для первого из них заключительное объединение происходит при показателе различия 80 (максимальное число в табл. 1), а для второго — при показателе различия 37.

3. При некоторых значениях расстояний, при которых происходят объединения, разбиения на класте-

ры двумя рассматриваемыми методами совпадают. Это имеет место при расстояниях 7 и 8 (строки 1 и 2 в табл. 2). В обеих дендрограммах первые объединения — попарные: в один кластер объединяются элементы 1 и 2 (расстояние между ними равно 7), а затем в другой кластер — элементы 15 и 16 (расстояние между ними равно 8).

4. Уже в строке 3 табл. 2, соответствующей расстоянию 14 между кластерами, проявляется отличие между двумя методами кластер-анализа. По методу «ближнего соседа» элемент 3 присоединяется к кластеру {1, 2}, поскольку $d(2, 3) = 14$, но при применении метода «дальнего соседа» такого не происходит, поскольку $\max[d(1, 3), d(2, 3)] = d(1, 3) = 22$. Элементы 1, 2, 3 объединяются в один кластер согласно методу «дальнего соседа» только при расстоянии 22 (строка 5 табл. 2). Это первое появление кластера более чем из двух элементов. Таким образом, элементы 1, 2, 3 образуют наиболее «сплоченную» тройку элементов при применении методов как «ближнего соседа», так и «дальнего соседа».

5. При дальнейшем объединении наблюдается естественное различие между классификациями, соответствующими двум методам. Метод «ближнего соседа» дает более крупные кластеры, чем метод «дальнего соседа», поскольку в первом случае присоединяемому элементу достаточно быть на определенном расстоянии от какого-либо одного элемента кластера, а во втором — расстояние до всех элементов кластера должно не превышать заданного значения. Тем не менее прослеживается некоторая «схожесть судеб» элементов и кластеров (типа той, что продемонстрирована выше в п. 4 на примере элементов 1, 2, 3).

6. Постепенно при увеличении расстояния, при котором объединяются кластеры, классификации по двум методам становятся все более похожими и, наконец, разбиение на 4 кластера — {1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 17}, {11, 12, 13}, {14, 15, 16} появляется в обеих дендрограммах (в табл. 2 строки 9 для метода «ближнего соседа» и 12 для метода «дальнего соседа»).

7. При следующем объединении видно различие: по методу «ближнего соседа» объединяются кластеры {10, 17} и {11, 12, 13}, а по методу «дальнего соседа» — кластеры {10, 17} и {14, 15, 16}.

8. Разбиение на два класса — одно и то же при применении обоих методов: {1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 11, 12, 13, 14, 15, 16, 17} (в табл. 2 строки 11 для метода «ближнего соседа» и 14 для метода «дальнего соседа»).

Таблица 2. Дендрограммы для методов «ближнего соседа» и «дальнего соседа»

Номер шага	Расстояние, при котором происходит объединение кластеров	Кластеры по методу «ближнего соседа»	Расстояние, при котором происходит объединение кластеров	Кластеры по методу «дальнего соседа»
1	7	{1, 2}, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17	7	{1, 2}, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17
2	8	{1, 2}, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, {15, 16}, 17	8	{1, 2}, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, {15, 16}, 17
3	14	{1, 2, 3}, 4, 5, 6, 7, 8, 9, 10, {11, 12}, 13, 14, {15, 16}, 17	14	{1, 2}, 3, 4, 5, 6, 7, 8, 9, 10, {11, 12}, 13, 14, {15, 16}, 17
4	18	{1, 2, 3}, 4, {5, 6}, 7, 8, 9, 10, {11, 12}, 13, 14, {15, 16}, 17	18	{1, 2}, 3, 4, {5, 6}, 7, 8, 9, 10, {11, 12}, 13, 14, {15, 16}, 17
5	19	{1, 2, 3, 4}, {5, 6, 7}, 8, 9, 10, {11, 12}, 13, 14, {15, 16}, 17	22	{1, 2, 3}, 4, {5, 6}, 7, 8, 9, 10, {11, 12}, 13, 14, {15, 16}, 17
6	20	{1, 2, 3, 4, 5, 6, 7}, 8, 9, 10, {11, 12}, 13, 14, {15, 16}, 17	25	{1, 2, 3}, 4, {5, 6}, {7, 8}, 9, 10, {11, 12}, 13, 14, {15, 16}, 17
7	24	{1, 2, 3, 4, 5, 6, 7}, 8, 9, 10, {11, 12}, 13, {14, 15, 16}, 17	32	{1, 2, 3}, 4, {5, 6}, {7, 8}, 9, {10, 17}, {11, 12}, 13, 14, {15, 16}
8	25	{1, 2, 3, 4, 5, 6, 7, 8, 9}, 10, {11, 12}, 13, {14, 15, 16}, 17	37	{1, 2, 3}, {4, 7, 8}, {5, 6}, 9, {10, 17}, {11, 12}, 13, 14, {15, 16}
9	32	{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 17}, {11, 12, 13}, {14, 15, 16}	38	{1, 2, 3, 9}, {4, 7, 8}, {5, 6}, {10, 17}, {11, 12}, 13, {14, 15, 16}
10	35	{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 11, 12, 13, 17}, {14, 15, 16}	43	{1, 2, 3, 9}, {4, 7, 8}, {5, 6}, {10, 17}, {11, 12, 13}, {14, 15, 16}
11	36	{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 11, 12, 13, 14, 15, 16, 17}	50	{1, 2, 3, 4, 7, 8, 9}}, {5, 6}, {10, 17}, {11, 12, 13}, {14, 15, 16}
12	37	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17}	60	{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 17}, {11, 12, 13}, {14, 15, 16}
13			66	{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 14, 15, 16, 17}, {11, 12, 13}
14			72	{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 11, 12, 13, 14, 15, 16, 17}
15			80	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17}

Таким образом, одновременное применение двух методов кластер-анализа дает возможность выделить классификации, совпадающие для обоих методов. Это разбиения на 2 и на 4 кластера:

{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 11, 12, 13, 14, 15, 16, 17},
{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 17},
{11, 12, 13}, {14, 15, 16}.

Именно эти два разбиения следует рассматривать как решение задачи кластер-анализа для данных табл. 1. Выбор одного из них проводится в соответствии с решаемой пользователем прикладной задачей.

Хотя разбиения, описанные в строках 1 и 2 табл. 2, также совпадают, их нецелесообразно рассматривать как решение задачи кластер-анализа, поскольку в них все кластеры, кроме одного (строка 1) или двух (строка 2), содержат единственный элемент.

Алгоритм расчетов для общего случая таков.

1. Построить две последовательности разбиений (две дендрограммы) методом «ближнего соседа» и методом «дальнего соседа».

2. Рассчитать расстояния между разбиениями для всех пар разбиений, одно из которых относится к первой последовательности, другое — ко второй. В качестве расстояния между разбиениями можно использовать расстояние Кемени или метрику подобия [8].

3. Найти минимум из указанных в п. 2 расстояний между разбиениями. Рассмотреть те пары разбиений, на которых минимум достигается. Очевидно, что для данных табл. 2 такими будут две пары совпадающих между собой разбиений на 2 и 4 кластера, которые приведены выше, поскольку любое расстояние между совпадающими распределениями равно нулю.

4. Если указанный в п. 3 минимум равен нулю, то пары разбиений, на которых минимум достигается, являются решением задачи кластер-анализа (именно их надо использовать из всех разбиений, содержащихся в дендрограммах). Если минимум не равен нулю, то выделяют устойчивые ядра разбиений. Для кластера A одного разбиения подбирают кластер B другого так, чтобы в их пересечении $A \cap B$ было максимальное число элементов. Тогда $A \cap B$ — устойчивое ядро. Перебирая все кластеры разбиения, получаем совокупность устойчивых ядер. При другом подходе к выделению устойчивых ядер рассматривают все пересечения $A \cap B$, в которых кластер A — из одного разбиения, кластер B — из другого, и выбирают те из них, в которых число элементов больше заданного пользователем порога. Совокупность устойчивых ядер разбиений отражает то общее, что имеется в двух классификациях. Принадлежность элементов классифицируемой совокупности, не вошедших в устойчивые ядра, к тому или иному кластеру зависит от выбора алгоритма классификации, т.е. она не является устойчивой. Решением задачи кластер-анализа следует считать совокупность устойчивых ядер разбиения с дополнительным указанием на остальные элементы как на информационный шум.

Пример. Рассмотрим два разбиения:
{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 11, 12, 13, 17}, {14, 15, 16};
{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10,14,15,16,17}, {11, 12, 13}.

Непустые попарные пересечения — это

{1, 2, 3, 4, 5, 6, 7, 8, 9}, {10, 17},
{11, 12, 13}, {14, 15, 16}.

При первом подходе кластеру {1, 2, 3, 4, 5, 6, 7, 8, 9} первого разбиения соответствует тот же кластер второго разбиения и первое устойчивое ядро — это {1, 2, 3, 4, 5, 6, 7, 8, 9}; кластеру первого разбиения {10, 11, 12, 13, 17} — кластер {11, 12, 13} второго разбиения и второе устойчивое ядро — это {11, 12, 13}; кластеру {14, 15, 16} — кластер {10, 14, 15, 16, 17} и третье устойчивое ядро — это {14, 15, 16}. В результате выделено три устойчивых ядра — {1, 2, 3, 4, 5, 6, 7, 8, 9}, {11, 12, 13}, {14, 15, 16}. Элементы 10 и 17 не вошли в устойчивые ядра, поскольку в зависимости от выбора алгоритма кластер-анализа они «перескакивают» от одного устойчивого ядра к другому.

При втором подходе результат зависит от порога. Если порог равен 1, то все непустые попарные пересечения являются устойчивыми ядрами; равен 2, то результат тот же, что и при первом подходе; равен 3, 4, 5, 6, 7, 8, то будут отброшены все кандидаты в устойчивые ядра, кроме единственного кластера из 9 элементов.

Итак, на основе концепции устойчивости предложен метод кластер-анализа, состоящий в построении дендрограмм методами «ближнего соседа» и «дальнего соседа» и выделении из них устойчивого разбиения. Такое разбиение дает более обоснованный результат по сравнению с применением одного из рассмотренных методов.

ЛИТЕРАТУРА

1. **Орлов А. И.** Устойчивость в социально-экономических моделях. — М.: Наука, 1979. — 296 с.
2. **Орлов А. И.** Устойчивые экономико-математические методы и модели. Разработка и развитие устойчивых экономико-математических методов и моделей для модернизации управления предприятиями. — Saarbrücken: Lambert Academic Publishing, 2011. — 436 s.
3. **Орлов А. И.** Устойчивые математические методы и модели / Заводская лаборатория. Диагностика материалов. 2010. Т. 76. № 3. С. 59 – 67.
4. **Дюран Б., Оделл П.** Кластерный анализ / Пер. с англ. — М.: Статистика, 1977. — 125 с.
5. **Дорофеюк А. А.** Алгоритмы автоматической классификации / Автоматика и телемеханика. 1971. № 12. С. 78 – 113.
6. **Мандель И. Д.** Кластерный анализ. — М.: Финансы и статистика, 1988. — 176 с.
7. **Орлов А. И.** О развитии математических методов теории классификации / Заводская лаборатория. Диагностика материалов. 2009. Т. 75. № 7. С. 51 – 63.
8. **Орлов А. И.** Прикладная статистика. — М.: Экзамен, 2006. — 671 с.