УДК 519.237

А.С. Герасимова

Кластеризация объектов с качественными признаками и ее использование для оценки силы их связи

A.S. Gerasimova

Clustering of Objects with Non-numerical Features and it's Use for the Strength of their Connections Estimation

Рассматривается задача применения к качественным категорированным данным классических алгоритмов кластеризации. Возникает задача присвоения категориям меток. Искомые метки должны быть наилучшим образом согласованы с совместными частотами встречаемости каждого из сочетаний категорий признаков. Предлагается вариант построения таких меток, и на основе него – способ оценивания силы взаимодействия нечисловых признаков.

Ключевые слова: кластеризация, нечисловые признаки, сила статистической связи.
DOI 10.14258/izvasu(2013)1.2-12

In the paper, we consider the problem of classical cluster algorithms applicability to the qualitative data. Assuming these data we face the problem of assigning the numerical labels to them. These labels must be coordinated with the observed joint frequencies of each combination of data categories with the best possible way. We propose a way to construct the labels. Also some estimation of the strength of non-numerical factors, based on constructed labels is proposed.

Key words: clusterization, non-numerical data, strength of statistical connection.

1. Вводные замечания. Постановка задачи. Под кластеризацией множества обычно понимают разбиение его элементов на группы в зависимости от их схожести. Сами группы принято называть кластерами. Кластерный анализ имеет большое количество применений во многих областях исследовательской деятельности.

В кластерном анализе для анализа количественных (числовых) данных имеется богатый арсенал статистических методов. Однако на практике (в частности, при обработке результатов медицинских или социологических исследований) часто приходится работать с данными, которые имеют качественный (нечисловой) характер. К ним нельзя применить многие классические методы статистической обработки, что, существенно затрудняя исследования, служит мотивацией для разработки способов работы с качественными данными.

Рассмотрим задачу кластеризации объектов с качественными категорированными признаками. Пусть при изучении n объектов $X_1, ...X_n$ с каждым из них связывается p качественных признаков $F_1, ...F_p$, причем i-й рассматриваемый признак F_i имеет m_i категорий.

Припишем категориям каждого из признаков числовые или векторные значения (метки). Как правило, сразу ясно, что замена названий или обозначений категорий просто их порядковыми номерами может привести к некорректному решению, так как присвоенные метки могут не отражать истинные различия между категориями. Понятие «истинного различия» здесь, конечно же, нуждается в уточнении.

Естественным (и, вероятно, единственно возможным) способом задания различий между качественными признаками является составление таблиц сопряженности признаков. Следовательно, вся доступная информация об истинных различиях категорий заключена в таблицах сопряженности. Это означает, что искомые метки должны быть наилучшим образом согласованы с совместными частотами встречаемости каждого из сочетаний категорий признаков. Степень согласованности найденных меток с такой таблицей сопряженности, таким образом, можно считать показателем качества получающегося решения.

Метки, найденные из условия максимизации описанного показателя, назовем частотно-согласованными. Если мы научились присваивать категориям признаков такие метки, то далее сможем построить кластерное разбиение наших данных одним из известных методов кластерного анализа. Предположим, что такое разбиение построено.

Пусть мы также располагаем некоторой априорной информацией об исходных данных (это может быть, например, пол испытуемых, социаль-

ная группа или предварительный диагноз у каждого из них). Тот показатель, который содержит в себе такую априорную информацию, предполагается также качественным категорированным. Будем назвать его объективным. Соответственно, мы имеем некоторое априорное разбиение данных на группы, соответствующие категориям объективного показателя, которое, конечно же, можно рассматривать как кластерное. Назовем такое разбиение объективным.

Если мы привлечем какой-либо метод сравнения между собой двух кластерных разбиений, то по степени различия объективного разбиения и разбиения, построенного нами, можно будет сделать выводы о степени связи качественных признаков с тем показателем, по которому строится объективное разбиение. Конкретно, чем более похожи эти разбиения, тем сильнее связаны наши признаки с объективным показателем.

Итак, перед нами стоят следующие задачи:

1. Присвоить категориям качественных признаков частотно-согласованные цифровые или векторные метки.

2. Найти и использовать какой-либо способ оценивать различия двух кластерных разбиений числом.

3. Предложить способ оценки силы влияния качественных признаков на некоторые «объективные» способы группировки исходных объектов, и, как следствие, способ оценивания силы взаимодействия каждого качественного категорированного признака и объективного признака (своеобразный коэффициент корреляции между ними).

2. Присвоение частотно-согласованных меток. Способ построения меток, согласованных с таблицами сопряженности, известен. Такой подход изучен довольно подробно в работах ряда французских статистиков и в современной интерпретации получил название анализа соответствий [1]. Его методики реализованы в виде отдельных блоков в большинстве статистических компьютерных пакетов. Например, анализ соответствий реализован в таких популярных статистических пакетах, как SAS, PASW (SPSS), STATISTICA.

В результате работы анализа соответствий каждая из категорий признаков может получить векторную метку размерности до m_i включительно (подробности можно найти в [2]). Поскольку координаты векторных меток формируются в порядке степени их разброса, то мы выберем в качестве числовых меток первые координаты получающихся векторных, как наиболее информативные.

Итак, каждый объект будет задан набором p чисел. Таким образом, мы приходим к стандартной задаче кластерного анализа. По полученным меткам построим кластерное разбиение с помощью какого-нибудь известного алгоритма кластеризации, например, с помощью алгоритма k-средних [2]. В случае $p = 2$ для практического решения поставленной задачи была написана компьютерная программа «Corr_an» на языке Delphi 7.0. Эта программа, кроме определения числовых меток категорий по заданной таблице сопряженности, находит также и их двумерные метки и строит рисунок.

3. Оценка различия двух кластерных разбиений. Чтобы сравнивать два кластерных разбиения будем использовать подход, предложенный в [3]. Коротко он может быть изложен так. Определим расстояние d на множестве всевозможных кластерных разбиений конечного множества $X = \{X_1, ...X_n\}$ формулой

$$d(A, B) \;=\; |A \triangle B| \;=\; |A \setminus B| + |B \setminus A|, \qquad (1)$$

где число элементов множества A обозначено как $|A|$. Тогда значение расстояния между кластерными разбиениями $\hat{A}$ и $\hat{B}$ можно задать формулой

$$d(\hat{A}, \hat{B}) \;= \sum_{x \in X} d(A_x, B_x), \qquad (2)$$

где $x \in A_x \in \hat{A},\ x \in B_x \in \hat{B}$.

Числовой коэффициент, называемый коэффициентом кластерных различий, определяется как

$$k(\hat{A}, \hat{B}) = 1 - \frac{d(\hat{A}, \hat{B})}{n(n-1)}. \qquad (3)$$

Он принимает значения от 0 до 1, и чем больше он по величине, тем более похожими друг на друга являются кластерные разбиения. Будем характеризовать степень различия разбиений с помощью этого коэффициента.

Вычислить коэффициент k можно с использованием специально написанной для этого компьютерной программы «ClasterRazb».

Построим кластерное разбиение исходного множества объектов, используя только значения одного из формирующих признаков, временно игнорируя значения остальных. Такое разбиение мы назовем индуцированным использованным признаком. Рассчитаем коэффициент кластерного различия индуцированного и объективного разбиений. Он будет тем больше, чем более похожи эти разбиения, а значит, чем более похожи эти два признака. В этом контексте он может рассматриваться как коэффициент кластерного сходства признаков.

Можно считать, что соответствующий коэффициент оценивает силу взаимодействия качественного признака и объективного. Если рассматривать в качестве двух разбиений те, которые индуцированы любыми из имевшихся признаков, то изложенный подход позволяет оценить силу их

связи. Используя в качестве индуцирующих сразу несколько признаков и оценивая различия разбиений, мы получим вариант коэффициента множественной корреляции между ними. Описанная методика с очевидными изменениями позволяет включить в рассмотрение также и числовые признаки и оценивать силу их связи.

4. Один практический пример. С целью опробовать описанную выше методику на практике были взяты данные исследования взаимосвязи генотипа человека с наличием у него повышенного риска образования венозных тромбов. Медицинские данные были предоставлены А.С. Петриковым, ангиохирургом отделения сосудистой хирургии МУЗ Городская больница №5 г. Барнаула.

Были обследованы 186 пациентов с 4 качественными генетическими признаками (аллелями FBG, PAI-1, GPIIB/IIIA и MTHFR) с 3 категориями каждый (нормозигота, гетерозигота, гомозигота).

Поскольку речь идет о наличии связи между генотипом пациента и риском развития венозного тромбоза, в качестве «объективного» разбиения на кластеры были выбраны две группы пациентов. Первая группа состояла из тех пациентов, которые страдали тромбозами в течение периода наблюдения. Остальные пациенты составляли вторую группу.

Итак, мы изучали различия в разбиении 186 объектов на два кластера при разных наборах качественных признаков. После оцифровки данных методом анализа соответствий каждая из категорий признаков получила частотно-согласованную векторную метку. Путем выбора определены цифровые метки первой координаты (см. табл. 1).

Таблица 1

Цифровые метки

	F_1	F_2	F_3	F_4
Нормозигота	0,002	0,220	-0,001	0,690
Гетерозигота	-0,04	-0,23	-0,10	-1,45
Гомозигота	0,155	-0,07	0,476	0,690

Таким образом, каждый из изучаемых объектов стал задаваться четырьмя числовыми показателями. Применив к полученным данным классический алгоритм k-средних, рассчитано разбиение множества данных на два кластера.

С помощью алгоритма оценки различий кластерных разбиений, описанного в п. 3, было произведено сравнение полученного разбиения с «объективным» разбиением. Коэффициент кластерного различия равен 0,506. Исходя из величины коэффициента, можно прийти к выводу об умеренной степени связи между генотипом пациента и наличием у него повышенного риска образования венозных тромбов.

Используем алгоритм, описанный в п. 4 для поиска возможно тесно связанных признаков. Построим кластерные разбиения множества объектов при удалении из множества признаков каждого из них и всевозможных их объединений. Найден коэффициент кластерного различия между построенными разбиениями, разбиением, полученным при учитывании влияния всех признаков, и объективным (см. табл. 2).

Таблица 2

Коэффициент кластерного различия с объективным разбиением

F	k	F	k
1	0,5031	2,3	0,5010
2	0,5031	2,4	0,5057
3	0,4982	3,4	0,5071
4	0,4994	2,3,4	0,4975
1,2	0,5043	1,3,4	0,5010
1,3	0,4978	1,2,4	0,5031
1,4	0,4975	1,2,3	0,5010

В столбцах, обозначенных F, собраны индуцирующие признаки.

Коэффициенты в этой таблице указывают на наличие умеренной силы связи всех комбинаций показателей с объективным показателем.

С помощью коэффициента кластерных различий оценим также силу попарных взаимодействий между качественными категорированными признаками относительно кластерной структуры множества.

Индуцируя кластерные разбиения каждым из признаков, найдем коэффициенты различий получающихся разбиений, после чего заполним таблицу 3.

Таблица 3

Коэффициенты кластерного сходства

Признак	1	2	3	4
1	1	0,69	0,82	0,51
2		1	0,64	0,52
3			1	0,51
4				1

Из таблицы 3 видно, что связи между первым и третьим, первым и вторым признаками можно считать сильными, остальные следует признать умеренными.

Поскольку, таким образом, есть статистически значимые связи между всеми имеющимися признаками, причем они однонаправлены, то все они имеют сходное воздействие на кластерную структуру множества объектов. Это позволяет отчасти объяснить примерно одинаковые величины коэффициентов в таблице 2.

Библиографический список

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. – М., 1989.

2. Дронов С.В. Многомерный статистический анализ: учебное пособие. – Барнаул, 2006.

3. Дронов С.В. Одна кластерная метрика и устойчивость кластерных алгоритмов // Известия АлтГУ. – 2011. – №1/2 (69).