УДК 519.254

Латентный кластерный анализ для случая двух кластеров

С.В. Дронов, А.Ю. Шеларь
АлтГУ, г. Барнаул

Ученый-исследователь в любой области знания обязательно сталкивается с различными множествами объектов и их классификаций. Однако, не всегда возможно однозначно определить класс объекта в случае, когда описания классов лишены полной четкости. Например, производится классификация по шкалам «высокий – низкий», «быстрый – медленный» и т. п., где каждая шкала разбивается на несколько групп. Возникающие группы в силу невозможности их строгого определения будем называть нечеткими категориями. Четким вариантом подобной нечёткой классификации является однозначное отнесение объекта к той или иной категории. При этом четкий выбор категорий можно интерпретировать, как придание конкретного числового выражения некоторому внешнему критерию оптимальности классификации, который раньше мог носить только интуитивный, непосредственно не наблюдаемый характер. Поэтому подобные методы сегодня принято называть методами латентной кластеризации.

Поставим задачу. Пусть есть множество объектов, для каждого из которых известны значения двух числовых признаков. Предполагается, что между этими признаками имеется зависимость. Требуется построить такое разбиение множеств значений каждого из признаков на классы, каждый из которых является четким вариантом одной из нечетких категорий, чтобы таблица сопряженности признаков в наибольшей степени отражала зависимость между ними. Таким образом, в роли внешнего критерия выступает степень зависимости изучаемых признаков.

Рассмотрим вариант задачи, в которой для каждого из признаков требуется построить лишь две категории. Например, можно рассматривать X – стаж работника, Y – размер его заработной платы. Предполагая наличие связи между этими признаками, исследователь хочет произвести разбиение значений X на категории «малый» и «большой» стаж, а Y на «низкая» и «высокая» заработная плата так, чтобы эта зависимость стала видна наиболее ясно.

Входными данными для проведения анализа будут являться результаты наблюдений над объектами, у каждого из которых известны значения X,Y. После решения этой задачи конкретные числовые значения признаков предполагается заменить их категориями, тем самым

превратив их в категорированные. При переходе к категориям в нашем случае все значения каждого из признаков, которые меньше соответственной границы, мы заменим на 0, остальные – на 1.

Обозначив за $a_{i,j}, i, j = 0,1$ количества объектов, у которых X принял значение i, а Y, соответственно, j, получим после выбора конкретных категорий четырехпольную таблицу с элементами $a_{i,j}$.

Нетрудно заметить, что при изменении границ, разбивающих множества значений признаков, будет изменяться и сама таблица. Зависимость признаков означает, что известная категория по X гарантирует известную категорию Y, а, следовательно, таблица окажется диагональной. Итак, разбиение множеств значений признаков будет тем более оптимальным, чем больше таблица сопряжённости похожа на диагональную.

Пусть N – количество изучаемых объектов. Имеем оценку вероятности F_{ij} попадания объекта в i, j клетку таблицы

$$F_{ij} = a_{ij} / N, \quad i, j = 0,1.$$

Если $F_{i,\bullet} = F_{i,0} + F_{i,1}$, $F_{\bullet,j} = F_{0,j} + F_{1,j}$, то, в случае независимости признаков, ожидаемые ее элементы будут вычисляться по формулам

$$E_{i,j} = N \cdot F_{i,\bullet} \cdot F_{\bullet,j}, \quad i, j = 0,1.$$

Чтобы таблица была похожа на диагональную, все так называемые смещения $S_{ij} = F_{ij} - E_{ij}$ должны быть сделаны максимально возможно большими. Справедливы утверждения

Лемма. *Величина* $|S_{00}|$ *однозначно определяет все* $|S_{ij}|$.

Теорема. *Для произвольных наборов значений двух признаков и любом определении границ категорий максимальное значение* $|S_{00}|$ *равно*

$$S_{\max} = \frac{1}{N^2} \cdot \left[\frac{N^2}{4} \right].$$

При этом оно достигается лишь на такой таблице, у которой вне одной из диагоналей расположены нули, а два оставшихся элемента одинаковы.

Согласно лемме, максимально возможное значение $|S_{00}|$ соответствует самой сильной степени связи между формирующими показателями. Следовательно, можно ввести коэффициент, значение которого позволяет численно оценить силу этой связи (по данным N выборочным значениям, [.] – целая часть числа):

$$\kappa = \frac{N^2 S_{00}}{\left\lfloor N^2 / 4 \right\rfloor}.$$

Ростовцевым [1] предложен алгоритм, позволяющий, по словам автора, построить такой вариант четких категорий каждого из признаков, которое позволяет достичь максимально возможного значения $|S_{00}|$ на изучаемом наборе данных. Однако нам удалось обнаружить, что это утверждение не вполне соответствует действительности.

Во-первых, для некоторых наборов данных величина $|S_{00}|$, значение которой приведено в теореме, не достигается после применения алгоритма Ростовцева. Во-вторых, то разбиение на категории, которое находит этот алгоритм, не является естественным даже в том смысле, на который указывает постановка задачи.

Действительно, пусть на значениях *X* и *Y* есть естественный порядок. Тогда образуемые категории должны тяготеть к противоположным концам шкал, т.е. значения каждого из показателей, близкие к началу шкалы должны относиться к 0-категории, а близкие к ее концу – к 1-категории. На подобный ожидаемый результат указывали и все примеры, приведенные в статье [1].

Но внекоторых практических примерах работы этого алгоритма, оказывается, что это далеко не так. Категории по *X* могут получаться «разрывными», что не отвечает постановке задачи.

Справедливости ради следует отметить, что в алгоритме Ростовцева построение категорий по *Y* всегда получается именно в «естественном» виде – фактически, в алгоритме производится полный перебор таких возможностей.

Значит, когда решается именно задача поиска наилучшей монотонной зависимости, алгоритм нуждается в доработке. При этом найденное значение $|S_{00}|$ может оказаться ниже, чем у Ростовцева – это плата за «естественность», «неразрывность» категорий.

Перейдем к описанию предлагаемой модификации алгоритма. Пусть имеется разбиение, полученное с помощью алгоритма Ростовцева. Построим по множеству значений *X* индикаторный вектор $I_x = (I_1, ..., I_n)$, элементы которого принимают значение 0 в случае если *i*-й элемент множества *X* нулевой категории, иначе 1. Необходимо преобразовать индикаторный вектор так, чтобы 0 и 1 свободно не чередовались, а распределялись по противоположным “концам” вектора I_x с минимально возможным уменьшением S_{00}.

1. Пусть I_{x_i} – *i*-е значение в I_x, где $i = 1...n-1$. $I_x' = I_x$.

2. Если $I_{x_i} = I_{x_{i-1}}$, то $i = i + 1$,и переходим к шагу 1, иначе:
3. Заменяем все значения в I_x левее I_{x_i} на I_{x_0}, справа – на $1 - I_{x_0}$.
4. В категорию X с номером I_{x_0} относим объекты с начала шкалы до последнего измененного значения. Найдем S_{00}. Заменяем все элементы правее $I_{x_{i-1}}$ на элементы из I_x' правее $I_{x_{i-1}}'$.

5. Увеличилось ли $|S_{00}|$? Если да, то запоминаем его, $i = i + 1$. К шагу 1. Иначе конец алгоритма.

В докладе рассмотрены соответствующие примеры и представлена программа для ЭВМ, реализующая описанный алгоритм.

Библиографический список

1. Ростовцев П.С. Черно-белый анализ связи переменных // Социология: 4М (методология, методы, математические модели). – 1998. – №10.

УДК 519.237

Оценивание силы пост-кластерной связи между формирующими показателями

С.В. Дронов, К.А. Леонгардт
АлтГУ, г. Барнаул

Предположим, что в результате работы некоторого кластерного алгоритма или путем экспертных оценок множество из n объектов $A_i, i = 1,...,n,$ каждый из которых задан совокупностью своих формирующих показателей $x_i, i = 1,...,k,$ разбито на p кластеров. Полученное кластерное разбиение будем далее называть основным и считать, что оно является объективно правильным в рамках решаемого круга задач. Другими словами, мы полностью доверяем тому алгоритму или той экспертной группе, в результате работы которых было построено основное разбиение.

Каждому кластеру поставлено в соответствие некоторое число которое условимся называть его меткой. Обозначим этот (внешний для решаемой задачи) набор меток **A**. Будем считать, что набор меток **A** значимо связан с основным кластерным разбиением и назовем этот набор основным. Предположение о связи основного кластерного разбиения с основным набором меток всегда справедливо, если и разбие-