УДК 519.237.8

Показатель качества кластеризации данных с бинарными признаками

В.В. Журавлева, Е.М. Марьин
АГУ, г. Барнаул

Современные организации работают с очень большим объемом исходной информации, что усложняет её понимание и анализ, а, как следствие, затрудняет ее использование при решении задач прогнозирования, управления, распознавания и многих других.

Автоматическая классификация заключается в разделении множества объектов различной информации на подмножества по их сходству или различию в соответствии с принятыми методами, обеспечивающими систематизацию объектов классификации по определенным выбранным признакам (свойства, характеристики или параметры объектов) [1]. Для бинарных данных значения признаков разделяются на нули или единицы (отсутствие или присутствие того или иного признака в конкретном наблюдении). Подобный вид данных весьма актуален в таких областях как, например, исследования цепочек ДНК.

Цель данной работы – разработка оригинального показателя качества кластеризации бинарных данных.

При исследовании структуры данных с помощью кластерного анализа немаловажную роль играет оценка качества кластеризации. Показатель качества кластеризации позволяет судить о том, насколько качественно была проведена кластеризация конкретным алгоритмом, либо из нескольких вариантов разбиений данных автоматически выбрать наиболее «разумный» [2, 3]. Показатель качества кластеризации как любая мера должен обладать такими свойствами как симметричность, транзитивность, ассоциативность.

Чаще всего, суть показателя заключается в попарном сравнении получившихся кластеров, а, затем уже в сравнении их совокупности.

Опишем последовательность построения показателя качества кластеризации для бинарных данных. Пусть $f_i(k)$ и $f_j(k)$ – частости значения «единица» для k-го признака в i-ом и j-ом кластерах. Тогда показатель отличимости кластеров по k-ому признаку определим следующим образом:

$$F_{ij}^k = 1 - \left|f_i(k) - f_j(k)\right|, \qquad (1)$$

где k – номер признака, i, j – номера кластеров. Итоговый показатель различимости двух кластеров определим как среднее геометрическое отличий по всем признакам:

$$Pr_{ij} = \sqrt[K]{F_{ij}^1 * \dots * F_{ij}^k}. \quad (2)$$

Интегральный показатель качества кластеризации будет иметь следующий вид:

$$T = \sqrt[M]{\prod_{\substack{i=1 \\ j>i}}^{k}(1 - Pr_{ij})}, \quad (3)$$

где $M = \frac{1}{2}(k^2 - k)$ – количество пар кластеров.

Для сравнения пар кластеров по отдельным признакам и взятия общего показателя различия используется формула (1). По значениям по парам признаков k, полученным в результате применения этой формулы можно говорить о том, что чем ближе результат к нулю, тем более явно различие между двумя признаками в кластерах. Результирующее значение ведёт себя аналогично результатам по парам признаков k из кластеров с номерами i, j.

При помощи формулы (2) формируется значение различности по всем возможным парам кластеров из получившегося результата. После чего мы получаем итоговое значение показателя качества кластеризации используя формулу (3). После применения формулы (3) о качестве проведённой кластеризации судят следующем образом: чем ближе итоговое значение к единице, тем лучше была проведена кластеризация и её объекты (кластеры) имеют ярко выраженное отличие друг от друга по тому или иному признаку, или же по ряду признаков.

Авторами разработан алгоритм расчёта описанного показателя качества кластеризации и проведено тестирование на модельных данных, которое показало, что идеи, заложенные в разработанный алгоритм оценки качества кластеризации, являются адекватными.

Библиографический список

1. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Института математики СО РАН., 1999. – 270 с.

2. Сивогловко Е.В. Оценка качества кластеризации в задачах интеллектуального анализа данных: Дис. … канд. физ.-мат. наук. – СПб. – 2014. – 92 с.

3. Журавлева В.В., Аюпов К.Е. О критериях оценки качества кластеризации // МАК: «Математики – Алтайскому краю»: сборник трудов всероссийской конференции по математике, Барнаул, 1–5 июля 2016 г. – Барнаул: Изд-во Алт. ун-та, 2016. – С. 130–131.