

3. Оскорбин Н.М., Боговиз А.В., Жариков А.В. Информационный аспект принятия решений в системе ЛПР // Динамика современной науки – 2011. Т. 2. Экономика : материалы VII международной научно-практической конференции. Республика Болгария, г. София, 17-25 июля 2011 г. – София: «Бял ГРАД-БГ» ООД, 2011. – С. 53–55.

4. Жариков А.В. Разработка математических моделей поддержки принятия решений при информационных ограничениях : дисс. ... канд. физ-мат наук. – Барнаул, 2011. – 122 с.

5. Петросян Л.А., Зенкевич Н.А., Семина Е.А. Теория игр : учебное пособие для университетов. – М. : Высш. шк., 1998. – 301с.

УДК 519.237

Работа с качественными нечисловыми данными при кластеризации объектов

А.С. Герасимова
АлтГУ, г. Барнаул

В практике анализа данных очень часто приходится работать с качественными (нечисловыми) данными. Как правило, к ним нельзя применить многие классические методы математической статистики, что, существенно затрудняя исследования, служит мотивацией для разработки новых способов работы с такими данными.

Рассмотрим задачу кластеризации объектов с качественными категоризованными признаками. Пусть при изучении n объектов X_1, \dots, X_n у них наблюдается p качественных признаков F_1, \dots, F_p , причем каждый рассматриваемый признак F_i имеет m_i категорий, $i=1, \dots, p$.

В работе рассматривается задача применения к таким данным классических алгоритмов кластеризации. Припишем каждой из категорий каждого из признаков числовое или векторное значение (метку). Процесс присвоения таких меток носит название оцифровки. Как правило, сразу ясно, что замена названий или обозначений категорий просто их порядковыми номерами может привести к некорректному решению – присвоенные метки могут не отражать истинные различия между категориями. Возникает задача присвоения категориям меток, согласованных с истинными различиями.

Естественный (и, вероятно, единственно возможный) способ задания различий между качественными признаками – составление таблиц сопряженности признаков. Это означает, что искомые метки должны быть наилучшим образом согласованы с совместными частотами

встречаемости каждого из сочетаний категорий признаков. Такой способ оцифровки 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 ряде популярных статистических пакетов, таких как SAS, PASW, STATISTICA.

В результате работы анализа соответствий каждая из категорий признаков может получить векторную метку размерности до m , включительно. Поскольку координаты векторных меток формируются в порядке степени их разброса, то мы выберем в качестве числовых меток первые координаты получающихся векторных, как наиболее информативные. Будем называть такие метки частотно-согласованными.

После замены категорий их частотно-согласованными метками каждый объект X_i , $i=1, \dots, n$ будет задан набором p чисел. Таким образом, мы приходим к стандартной (числовой) задаче кластерного анализа. Построим кластерное разбиение исходного набора объектов 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 «Cог_an» на языке Delphi 7.0, которая по таблице сопряженности признаков формирует двумерные векторные метки категорий обоих признаков и строит рисунок, изображающий эти категории точками в двумерной системе координат.

Описанная выше методика была опробована на практике с помощью данных исследования взаимосвязи генотипа человека с наличием у него повышенного риска образования венозных тромбов. Реальные медицинские данные предоставлены Петриковым А.С., ангиохирургом отделения сосудистой хирургии МУЗ Городская больница №5 г. Барнаула.

Были обследованы 186 пациентов с 4 качественными генетическими признаками (аллелями FVG, PAI-1, GPIIb/IIIa, MTHFR) с 3 категориями каждый (нормозигота, гетерозигота, гомозигота). Таким образом, мы изучали различия в разбиении 186 объектов на два кластера («здоров»/«болен») при разных наборах качественных признаков. После оцифровки данных методом анализа соответствий каждая из категорий признаков получила частотно-согласованную векторную метку. Путем выбора в качестве цифровой метки первой координаты получили цифровые метки объектов (табл. 1).

Каждый из изучаемых объектов стал задаваться четырьмя числовыми показателями. После применения к полученным данным класси-

ческого алгоритма k -средних, было получено разбиение исходного множества данных на 2 кластера.

Таблица 1

Цифровые метки объектов				
	F_1	F_2	F_3	F_4
нормозигота	0,002	0,220	-0,001	0,690
гетерозигота	-0,043	-0,227	-0,101	-1,449
гомозигота	0,155	-0,065	0,476	0,690

Полученный результат сравнивался с объективно правильным разбиением, построенным на основе клинического обследования, и показал высокие чувствительность (81,43%) и специфичность (90,95%).

Библиографический список

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.
2. Дронов С.В. Многомерный статистический анализ: учебное пособие. – Барнаул: Изд-во Алт. ун-та, 2006. – 213 с.

УДК 517.11:518.5

Модель машины Шенфилда с оракулом и ограничением на время

В.Р. Карымов
АлтГУ, г. Барнаул

Рассматриваются машины Шенфилда с оракулом, которые имеют потенциально бесконечное число регистров, управляющее устройство и программу. Регистры пронумерованы натуральными числами, и каждый из них может содержать натуральное число, букву или некоторый символ. Пусть x_n обозначает содержимое регистра с номером n . Как станет ясно из дальнейшего, для любой машины необходимо лишь конечное множество регистров и конечное множество используемых символов, и эти множества нетрудно заранее определить по программе исходной машины.

В каждой машине выделены специальные регистры для записи технической информации. Пусть в регистрах 0, 1 расположены счетчик команд и счетчик тактов; в регистрах \bar{p} и \bar{z} записаны программа рас-