УДК 519.25

Иерархическая оцифровка в задачах однородности

С.С. Никеев, С.В. Дронов
АлтГУ, г. Барнаул

В работе рассматривается проблема однородности данных. Классически проблема однородности в ее простейшем варианте выглядит так (см., например, [1]): даны два набора объектов. Следует ли считать, что объекты одного из них не имеют существенных различий по отношению к объектам другого? Есть и еще одна практически важная задача, близкая к задаче однородности. Пусть имеется множество X объектов. Требуется понять, существует ли естественное расслоение его на кластеры, т.е. подмножества элементов X, более близких друг к другу, чем к элементам других получающихся подмножеств. Эта задача часто встречается в приложениях как вопрос о необходимости стратификации изучаемого набора объектов в противовес изучению его в целом. Ее с нашей точки зрения тоже можно считать проблемой однородности.

Рассмотрим новый подход к задаче однородности, основанный на применении к ней методов кластерного анализа. Будем считать, что у нас имеется самый простой случай, когда имеются два класса объектов, причем каждый объект полностью характеризуется одним параметром, и, следовательно, может изображаться точкой на действительной прямой (размерность задачи равна 1). Пусть в первом классе у нас имеется n_1 элементов, а во втором классе – n_2 элементов. Всего, таким образом, имеется $n = n_1 + n_2$ объектов.

Объединим два имеющихся у нас класса в универсальное множество X и применим к этому множеству иерархический алгоритм кластерного анализа в его агломеративном варианте (разработка [2]).

В процессе работы алгоритма все изначально «рассыпанные» объекты из X последовательно собираются группы близких к друг другу, на последнем шаге объединяясь в одну группу. На каждом шаге к какой-либо группе присоединяется тот объект, который ближе к этой группе, чем все оставшиеся. Это дает возможность предположить, что, если исходные классы объектов не были однородными, то объекты из разных классов долго не объединятся в одну группу во время работы алгоритма.

Заметим, что работу иерархического алгоритма можно остановить в любой момент, и группы, которые в этот момент уже сформированы, объявить кластерами. Имея в виду это, сделаем остановку, когда все элементы одного из классов соберутся в один кластер. Подсчитаем количество элементов другого класса в этом кластере. Очевидно, можно считать, что чем меньше это количество, тем классы однородны в меньшей степени.

Это позволяет ввести специальный коэффициент однородности, принимающий значения от 0 до 1. Пусть в момент, когда все элементы одного из классов впервые собрались в кластер, в нем оказалось m элементов другого класса. Тогда определим коэффициент кластерной однородности формулой

$$\Re = \frac{m}{n_2}.$$

Если он равен 1, то изучаемые классы будем считать однородными в полной степени. Если 0, то полностью неоднородными.

Простейший вариант решения задачи однородности: если при описанной выше процедуре получим $\Re = 1$, то классы признаются однородными, иначе нет. Этот вариант назовем жестким критерием. Предложенный способ можно обобщить: зададим малое положительное число ε. Если $\Re > 1 - \varepsilon$, то классы однородны на уровне ε. Такой способ принятия гипотезы однородности будем называть ε-мягким критерием.

Один из наиболее распространенных способов проверки гипотезы однородности связан с применением критерия Пирсона χ^2. Сравним результат предложенного выше жесткого критерия этой процедурой.

В ситуации, когда сработал жесткий критерий, объекты одного из классов ближе собраны к какому-то центру, чем для второго класса.

Лемма. *Пусть заданы два класса объектов* $X, Y \subset R$, *такие, что*

$$X = \{x_j, j = 1,\ldots,n_1\}, \quad Y = \{y_i, i = 1,\ldots,n_2\}$$

и, существует такой z, что

$$\forall i, j \quad \rho\left(x_j, z\right) < \rho\left(y_i, z\right), \tag{1}$$

где ρ –метрика в R. Тогда можно так образовать группы для χ^2*, что в каждой из них окажутся лишь элементы одного из классов.*

Доказательство. Рассмотрим $r = \max\limits_j \rho\left(z, x_j\right)$. Тогда из (1) следует, что в интервалах $[z-r, z] = \Delta_i$ и $[z, z+r] = \Delta_{i+1}$ не содержится не одного элемента из Y. Также в этих интервалах содержатся все элементы X, не содержится не одного элемента из Y. Также в этих интервалах содержатся все элементы X. Эти интервалы включим в число строящихся для работы критерия Пирсона, остальные интервалы можно строить произвольным образом. Лемма доказана.

Из этой леммы немедленно вытекает следующий результат.

Теорема. *Пусть заданы два класса объектов* $X, Y \subset R$ *с условием (1). Тогда можно разбить числовую прямую на интервалы так, что по этому разбиению критерий Пирсона отвергнет гипотезу однородности.*

При доказательстве теоремы проверяется, что на построенных в лемме интервалах статистика критерия Пирсона принимает свое теоретически максимально возможное значение $p_f = s_1 p_1 + s_2 p_2$ Эта теорема дает повод считать, что в описанной ситуации при построении интервалов для работы критерия хи-квадрат произвольным образом значение статистики критерия будет если и не максимально возможным, то близким к нему. Таким образом, по крайней мере, в случае ярко выраженной неоднородности классический и новый алгоритм дадут одинаковые результаты.

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. Андреева В.Н., Дронов С.В. Визуализация иерархических кластерных алгоритмов // Сборник трудов семнадцатой региональной конференции по математике «МАК-2014», посвященной 40-летию факультета математики и информационных технологий. – Барнаул: Изд-во Алт. ун-та, 2014. – С. 16–19.

УДК 514.182

Методы изображения геометрических фигур

Д.И. Оглезнев, И.В. Пономарев
АлтГУ, г. Барнаул

В процессе визуализации результатов решения большого числа задач, часто требуются изображения различных трехмерных геометрических тел. При этом исследователь сталкивается с проблемой наиболее наглядного представления получаемых тел на плоскости. Эта задача осложняется еще и тем, что не все графические компьютерные программы обладают возможностью представления трехмерных объектов.

Для изображения трехмерных геометрических тел на плоскости обычно используют параллельное или центральное проектирование. Задача заключается в том, чтобы по координатам точек оригинала (X, Y, Z) получить координаты точек изображения (x, y). В методе параллельных проекций используют следующую теорему [1].

Теорема 1. Координаты точки-изображения суть линейные функции координат точки-оригинала, т. е.

$$\begin{aligned} x &= a_1 X + b_1 Y + c_1 Z + d_1; \\ y &= a_2 X + b_2 Y + c_2 Z + d_2, \end{aligned} \tag{1}$$

где $a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2$ – некоторые постоянные коэффициенты. Согласно теореме Польке-Шварца [4], для построения однозначного изображения достаточно задать проекции четырех некомпланарных точек. Например, при ортогональном проектировании на плоскость $X + Y + Z = 0$ точки