Секция 7. BIGDATA–ТЕХНОЛОГИИ АНАЛИЗА И ПРОГНОЗИРОВАНИЯ

ИСПОЛЬЗОВАНИЕ НЕПАРАМЕТРИЧЕСКИХ СВОЙСТВ ПОРЯДКОВЫХ СТАТИСТИК В ЗАДАЧАХ DATA MINING

Кобзев В.Г.
Харьковский национальный университет радиоэлектроники

Одной из задач, рассматриваемых во многих исследованиях по Data Mining, например, в работе [1], является выявление аномальных значений (выбросов) в совокупностях анализируемых данных. При этом важно учитывать исходные знания (предположения) о характере их статистического распределения [2], а также количество значений, подлежащих анализу.

В [3] рассматривается задача проверки на аномальность числовых результатов последовательно проводимых экспериментов с нарастающим количеством наблюдений величин, обладающих общими статистическими свойствами. Для классического понимания аномальности на первом шаге производится частичное упорядочение имеющейся выборки результатов первого эксперимента и проверка ее экстремальных значений на аномальность с помощью одного из характерных критериев. Если последовательно появляющиеся результаты новых экспериментов являются экстремальными в образуемых выборках, то они подлежат проверке на аномальность с помощью ранее выбранного критерия с учетом вновь образовавшегося объема выборки.

После каждого эксперимента возможно более глубокое изучение соответствия структуры новой выборки анализируемых величин на основе наиболее правдоподобных границ порядковых статистик для предполагаемого распределения экспериментальных данных. Такими границами являются точки пересечения кривых плотностей соседних порядковых статистик в выборке объема n из совокупности с произвольным непрерывным распределением $F(x)$, которые согласно [2], являются квантилями этого распределения уровней i/n, $i=\overline{1,n-1}$. В качестве аномальных могут быть признаны как экстремальные значения, так и любые другие элементы анализируемой выборки. Следует отметить, что в указанные границы попадают математические ожидания порядковых статистик $F^{-1}(\frac{i}{n+1})$ при всех значениях $i=\overline{1,n-1}$. Нетрудно заметить, что они абсолютно точно совпадают с наиболее правдоподобными границами порядковых статистик для совокупности на единицу большего объема. Данный факт используется для анализа на аномальность выборок последовательно нарастающего объема.

1. Han J., Kamber M., Pei J. Data Mining. Concepts and Techniques. 3-d edition. – Elsevier, 2012. – 703p. 2. Кобзев В.Г. Обнаружение выбросов с использованием непараметрических свойств порядковых статистик / ИСТ-2016. науч.-техн. конф.: тезисы докладов. – Х.: ДРУКАРНЯ МАДРИД, 2016. – с. 321-322. 3. Kobziev V.G. Technology data analysis on the anomalous in a sequence of experiments / – Харьков: ННЦ ХФТИ, 2017, с. 52.