ОБНАРУЖЕНИЕ ВЫБРОСОВ С ИСПОЛЬЗОВАНИЕМ НЕПАРАМЕТРИЧЕСКИХ СВОЙСТВ ПОРЯДКОВЫХ СТАТИСТИК

Кобзев В.Г.

Харьковский национальный университет радиоэлектроники

Важной составной частью работ в любой сфере экспериментальных исследований являются анализ и статистическая обработка данных. Одной из проблем статистической обработки данных является определение аномальных значений (выбросов) в совокупностях результатов проведенных экспериментов.

Выбросом (аномальным значением) принято считать элемент совокупности, значительно отличающийся набором своих характеристик (значений) от остальных элементов. Термин «значительное отличие» не имеет однозначного толкования. В то же время, желательно иметь значение количественной меры соответствия одного или нескольких особенных элементов анализируемой выборки данных исходным предположениям о характере их статистического распределения. Эти предположения, как правило, сводятся к некоторым общим допущениям (непрерывность, унимодальность, симметрия) относительно плотности распределения значений изучаемой величины или более конкретным утверждениям, например, анализируемые данные подчинены нормальному распределению с неизвестными параметрами.

Существует несколько типов выбросов (глобальные, контекстуальные, коллективные), и группы подходов к их выявлению. Известные методы проверки (тесты) на наличие выбросов разделяются на группы с обучением, самообучением и без использования обучения.

В работе [1] приведены два достаточно известных статистических метода проверки одного подозреваемого значения на аномальность. Рассматривается выборка наблюдений некоторой случайной величины, предположительно подчиняющейся нормальному закону распределения $N(\mu,\sigma)$.

Первый метод основан на известном факте, заключающемся в том, что интервал $\mu \pm 3\sigma$ содержит 99,7% значений из их генеральной совокупности.

Второй метод использует несколько менее известный факт: в интервал

$$[\ Q_{1/4} - 1{,}5*(Q_{3/4} - Q_{1/4});\ \ Q_{3/4} + 1{,}5*(Q_{3/4} - Q_{1/4})\ ] \qquad (1)$$

где Q_p - квантиль уровня p статистического распределения изучаемой величины $P\{x \le Q_p\} = p$, в случае гауссова (нормального) распределения попадает 99,3% значений. Видно, что этот интервал образуют границы, на полтора межквартильного размаха удаленные в противоположные стороны от нижнего (влево) и верхнего (вправо) квартиля.

В обоих методах выбросом предписано признавать одно значение, лежащее вне указанных интервалов, так как его появление имеет слишком малую вероятность. Для определения граничных значений интервалов могут использоваться максимально правдоподобные или другие оценки параметров статистических распределений.

Предположение о гауссовом распределении анализируемых данных во многих случаях требует дополнительного подтверждения. Непараметрические

статистические методы не используют предположений о возможности описать плотность распределения анализируемых случайных величин с помощью конечного количества параметров.

Авторы работы [1] и других известных работ рекомендуют два метода выявления выбросов: 1) с использованием гистограмм с равными интервалами группировки и 2) с использованием ядерных оценок плотности распределения. В первом методе выбросом признается значение, находящееся в интервале с частотой попадания менее выбранной критической величины. Во втором, аналогично случаю с известным распределением, выброс – значение в области маловероятных значений, определенной по построенной оценке плотности.

Отметим, что ядерные функции используют влияние каждой выборочной точки на своих соседей. Взаимосвязь одномерных значений анализируемой совокупности $x_1, x_2, ..., x_j, ..., x_n$ отображает вариационный ряд

$$x_{(1)} \le x_{(2)} \le ... \le x_{(i)} \le ... \le x_{(n)}.$$

С целью исследования зон концентрации значений порядковых статистик автор рассмотрел области их наиболее правдоподобных значений [2]. Под областью наиболее правдоподобных значений *i*-й порядковой статистики понимается область таких значений случайной величины *X*, в которых функция плотности $\varphi_{i,n}(x)$ доминирует над функциями плотностей всех остальных порядковых статистик. Доказано [2], что абсциссами точек пересечения кривых плотностей соседних порядковых статистик в выборке объема n из совокупности с произвольным непрерывным распределением $F(x)$ являются квантили этого распределения уровней $i/n,\ i=\overline{1,n-1}.$

Последнее позволяет в зависимости от величины объема выборки n построить процедуру оценивания квантиля требуемого уровня. Квантили уровней 0,25 и 0,75 (квартили) могут быть использованы в описанном выше методе для определения критического интервала (1) и проверки подозреваемого значения на аномальность. Квантили уровней 0,05 и менее могут быть определены в выборках объемом от 20 значений и более, затем их можно непосредственно использовать для проверки наличия выбросов.

Приводятся результаты использования свойств порядковых статистик при анализе наличия выбросов в данных экспериментальных исследований [3].

1. Han J., Kamber M., Pei J. Data Mining. Concepts and Techniques. 3-d edition. – Elsevier, 2012. – 703p. 2. Кобзев В.Г. Исследование статистических свойств параметров однотипных элементов / IV Межд. научн. конф. «Функциональная база наноэлектроники». Сборник научных трудов – Харьков: ХНУРЭ, 2011. – с. 279-281. 3. Кобзев В.Г. Технология последовательного анализа экспериментальных данных на аномальность / Сб. материалов 13-й конф. по физике высоких энергий, ядерной физике и ускорителям. - Харьков, ННЦ ХФТИ, 2016. - с. 108.