Вестник КРАУНЦ. Физ.-мат. науки. 2018. № 3(23). С. 190-194. ISSN 2079-6641

DOI: 10.18454/2079-6641-2018-23-3-190-194

УДК 519.7

ПРИМЕНЕНИЕ НЕПРЕРЫВНОЙ ЛОГИКИ ДЛЯ ОПРЕДЕЛЕНИЯ ГРАНИЦ ВЫБРОСОВ

Е. В. Шматова

Институт прикладной математики и автоматизации – филиал ФГБНУ «Федеральный научный центр «Кабардино-Балкарский научный центр Российской академии наук» (ИПМА КБНЦ РАН) , 360000, г. Нальчик, ул. Шортанова, 89А

E-mail: lenavsh@yandex.ru

Аннотация. В настоящей работе предложен метод применения непрерывной логики для выявления выбрасов данных с целью избавления от шумов и ошибочных значений.

Ключевые слова: непрерывная логика, выбросы, интервал, классы.

MSC 68T27

APPLICATION OF CONTINUOUS LOGIC FOR DETERMINING EMISSION LIMITS

E. V. Shmatova

Institute of Applied Mathematics and Automation of Kabardin-Balkar Scientific Centre of RAS (IAMA KBSC RAS), 360000, Nalchik, Shortanova st., 89A

E-mail: lenavsh@yandex.ru

In this paper, we propose a method of applying continuous logic to identify data ejections in order to eliminate noise and erroneous values.

Key words: continuous logic, emissions, interval, classes.

Введение

Машинное обучение включает в себя методы построения моделей и алгоритмов, способных обучаться на данных. В современном мире данные, как правило, представляют собой огромные массивы информации, в которых необходимо выявить скрытые закономерности. При этом в массиве данных могут присутствовать ошибочные или неточные сведения об исследуемых процессах (объектах) [1]. В этой связи возникает задача фильтрации выбросов — это обнаружение в обучающей выборке небольшого числа нетипичных объектов. В некоторых задачах их поиск является целью, например, обнаружение мошенничества. В других - эти объекты являются следствием ошибок в данных или неточности модели, то есть шумом, мешающим настраивать модель, и должны быть удалены из выборки.

Проблема выбросов

Выбросы являются значениями, которые сильно отличаются от большинства типичных значений в наборе данных. При небольших объемах данных выбросы легко обнаруживаются в таблицах значений, или визуально на графиках. Однако, при наличии больших объемах исходных данных, когда визуальное определение выбросов становится мало эффективно, желательно опираться на общие методы выявления выбросов. Часто используются метод вычисления границ выбросов Тьюки [2]:

1) Необходимо упорядочить данные по возрастанию.

2) Вычисляется медиана набора данных Q2. Медиана набора данных – это величина, находящаяся в середине набора данных. Если набор данных содержит нечетное количество значений, то медиана – это значение, до которого и после которого расположено одинаковое количество значений в наборе данных. Но если набор данных содержит четное число значений, то нужно найти среднее арифметическое двух средних значений.

3) Вычисляется нижний квартиль Q1, ниже которой лежит 25 % значений из набора данных, т.е. это половина значений, расположенных до медианы. Если до медианы лежит четное количество значений из набора данных, нужно найти среднее арифметическое двух средних значений (это аналогично вычислению медианы).

4) Вычисляется верхний квартиль Q3, выше которой лежит 25 % значений из набора данных, т.е. это половина значений, расположенных после медианы. Процесс вычисления Q3 аналогичен процессу вычисления Q1.

5) Вычисляется межквартильный диапазон. Вычислив Q1 и Q3, необходимо найти расстояние между этими величинами. Для этого вычтите Q1 из Q3. Значение межквартильного диапазона крайне важно для определения границ значений, которые не являются выбросами.

6) Находятся «внутренние границы» значений в наборе данных. Выбросы определяются через анализ значений – попадают ли они или нет в пределы так называемых «внутренних границ» и «внешних границ». Значение, лежащее вне «внутренних границ», классифицируется как «незначительный выброс», в то время как значение, находящееся за «внешними границами», классифицируется как «значительный выброс». Чтобы найти внутренние границы, необходимо умножить

межквартильный диапазон на 1,5; результат нужно прибавить к Q3 и вычесть из Q1. Два найденных числа являются внутренними границами набора данных.

7) Находятся «внешнии границы» набора данных: межквартильный диапазон умножается на 3. Результат нужно прибавить к Q3 и вычесть из Q1. Два найденных числа являются внешними границами набора данных.

Этот метод позволяет определить, являются ли некоторые значения выбросами (незначительными или значительными). Тем не менее, значение, классифицируемое в качестве выброса, является только «кандидатом» на исключение, то есть не обязан быть исключенным. Некоторые выбросы должны быть исключены из набора данных, так как их причинами являются ошибки и технические неполадки; другие выбросы необходимо оставить в наборе данных. Если, например, выброс не является результатом ошибки и/или дает новое понимание тестируемого явления, то его необходимо оставить в наборе данных.

Границы выбрасов в рамках непрерывной логики

Для логического описания моделей с выбросами будем использовать непрерывную логику, методы и функции которой наглядно демонстрируют удобство и адекватность этого аппарата как средство построения логических процедур выявления выбросав в задачах интеллектуальной обработки данных.

Сформулируем основные свойства непрерывной логики [3].

Пусть $C = [A,B]$ замкнутый интервал с серединой $M = (A+B)/2$. Основные операции НЛ определяются на С:

– отрицание $\bar{a} = 2M - a$;

– дизъюнкция $a \vee b = max(a,b)$,

– конъюнкция $a \wedge b = min(a,b)$.

Алгебры, образуемые множеством C вместе с теми или иными базовыми операциями на нем, называются алгебрами НЛ. Любая функция вида $C^n \to C$, в форме суперпозиции конечного числа базовых операций данной алгебры НЛ, примененных к аргументам $x_1,...,x_n \in C$, называется функцией НЛ.

Непрерывная логика является непосредственным обобщением дискретной логики на случай непрерывного носителя C. Большинство законов дискретной логики сохраняется и в непрерывной логике. Только законы противоречия и исключенного третьего дискретной логики здесь не действуют, а заменяются на следующие:

$a\bar{a} = M - |a - M|$,

$a \vee \bar{a} = M + |a - M|$.

Рассмотрим вопрос о выявлении границ выбрасов в рамках непрерывной логики.

Пусть заданы признаки $X = \{X_i\}$, $X_i = (x_1^{(i)}, x_2^{(i)}, ..., x_n^{(i)}), i = 1,2,...,m$, характеризующие множество объектов $Z = \{Z_1, Z_2, ..., Z_m\}$, где $X \in [A,B]$, $x_j \in [a_j^{'}, b_j^{'}]$, $[a_j^{'}, b_j^{'}] \subset [A,B]$. Т.е.

$$X_1 \to Z_1$$
$$X_2 \to Z_2$$
$$...$$
$$X_m \to Z_m$$

Пусть определен некоторый объект Z_y характеризующийся набором свойств $Y = (y_1, y_2, ..., y_n)$, $y_j \in [c_j, d_j]$, $[c_j, d_j] \subset [A,B]$, $j = 1,2,...,n$.

Необходимо определить принадлежит ли объект Z_y к какому либо из ранее выделенных классов, образованных множеством объектов $Z_1,...,Z_n$. Иными словами требуется определить степень принадлежности объекта Z_y к какому либо из существующих классов, в противном случае объект Z_y будет классифицирован в качестве выбраса данных.

Если объект Z_y будет принадлежать одному из классов, то он будет добавлен во множество объектов Z в качестве $m+1$ элемента. В противном случае оказавшись выбросом данных объект Z_y будет отброшен с тем чтобы не вносить изменений (шумов) в ранее выявленную систему знаний.

Заметим, что выявленные классы $K = K_1,...,K_l$ определяются некоторым наборами характеристик из множества $\{x_1,....,x_n\}$. Каждый такой набор может быть записан в виде конъюнкций соответствующего этому классу набора характеристик.

Найдем растояние от $Y = (y_1,y_2,...,y_n)$ до $X_i = (x_1^{(i)},x_2^{(i)},...,x_n^{(i)}), i = 1,2,...,m$ покомпонентно. В простейшем случае определим расстояние между характеристиками x_j и y_j как расстояние между серединами соответствующих этим характеристикам интервалов. То есть

$m_j = \dfrac{a_j^{'}+b_j^{'}}{2}$ середины интервалов характеристик X,

$e_j = \dfrac{c_j+d_j}{2}$ середины интервалов характеристик Y.

Среди всех $m_j, j = 1,...,n$ определим ближайший к левой границе $[A,B]$ и обозначим через $m_{\text{лев}}$, а ближайший к правой границе $[A,B]$ обозначим $m_{\text{пр}}$.

Если большинство e_j содержатся вне интервала $[m_{\text{лев}},m_{\text{пр}}]$, то Y можно классифицировать в качестве выброса.

Отсюда видно что для анализа принадлежности объекта Z_j к фиксированному классу $K_1,...,K_l$ необходимо провести описанное выше покомпонентное сравнение характеристик Y с теми значениями X_i которые и образуют рассматриваемый класс составляя соответствующую ему конъюкцию. В результате покомпонентного сравнения объект не содержащийся внутри интерва ни одного из выделенных классов $K_1,...,K_l$ является выбросом.

Работа выполнена при поддержки гранта РФФИ № 18-01-00050-а.

Список литературы

[1] Дьяконов А. Г., Головина А. М., “Выявление аномалий в работе механизмов методами машинного обучения”, *Аналитика и управление данными в областях с интенсивным использованием данных*, труды XIX Международной конференции DAMDID/RCDL’2017, 2017, 469–476. [D’yakonov A. G., Golovina A. M., “Vyyavlenie anomalij v rabote mekhanizmov metodami mashinnogo obucheniya”, *Analitika i upravlenie dannymi v oblastyah s intensivnym ispol’zovaniem dannyh*, trudy XIX Mezhdunarodnoj konferencii DAMDID/RCDL’2017, 2017, 469–476].

[2] Форман Дж., “Много цифр: Анализ больших данных при помощи Excel”, 2016, 461 с. [Forman Dzh., “Mnogo cifr: Analiz bol’shih dannyh pri pomoshchi Excel”, 2016, 461 pp.]

[3] Левин В. И., “Логико-математические методы и их применения”, *Системы управления, связи и безопасности*, **2** (2018), 213-244. [Levin V. I., “Logiko-matematicheskie metody i ih primeneniya”, *Sistemy upravleniya, svyazi i bezopasnosti*, **2** (2018), 213-244].

Список литературы (ГОСТ)

[1] Дьяконов А. Г., Головина А. М. Выявление аномалий в работе механизмов методами машинного обучения // Аналитика и управление данными в областях с интенсивным использованием данных: труды XIX Международной конференции DAMDID/RCDL'2017. 2017. С. 469–476.
[2] Форман Дж. Много цифр: Анализ больших данных при помощи Excel. М.: Альпина Паблишер, 2016. 461 с.
[3] Левин В. И. Логико-математические методы и их применения // Системы управления, связи и безопасности. 2018. №. 2. С. 213-244.

Для цитирования: Шматова Е. В. Применение непрерывной логики для определения границ выбросов // *Вестник КРАУНЦ. Физ.-мат. науки*. 2018. № 3(23). С. 190-194. DOI: 10.18454/2079-6641-2018-23-3-190-194

For citation: Shmatova E. V. The application of continuous logic to determine emission limits, *Vestnik KRAUNC. Fiz.-mat. nauki*. 2018, **23**: 3, 190-194. DOI: 10.18454/2079-6641-2018-23-3-190-194

Поступила в редакцию / Original article submitted: 08.06.2018