УДК 519.7 doi: 10.26102/2310-6018/2018.23.4.011

Л.А. Лютикова

# ПОСТРОЕНИЕ ЛОГИЧЕСКОГО АЛГОРИТМА ВЫЯВЛЕНИЯ ВЫБРОСОВ В ЗАШУМЛЕННЫХ ДАННЫХ

*Институт прикладной математики и автоматизации – филиал Федерального государственного бюджетного научного учреждения «Федеральный научный центр «Кабардино-Балкарский научный центр Российской академии наук», Нальчик, КБР*

*В работе предложен логический подход к анализу качества данных для решения задач машинного обучения. При разработке алгоритмов машинного обучения часть исходных данных решаемой задачи объединяют в обучающую выборку. Как правило, качество этих данных не вляется идиальным, и это дастаточно острая проблема возникающая при построении обучающих систем распознавания. Так как построение модели распознования является результатом последовательного предъявления исходного набора данных, то их некорректность может существенно искозить конечную модель, что скакжется на результатах работы алгоритмов распознования. Данные, которые вносят искажения при построении модели называют выбросами. Причиной возникновения выбросов являются помехи аппаратуры, неверная интерпритация эксперта, шумы и т.д. В связи с этим возникает задача анализа данных на предмет выявления выбросов и ослобления их влияния на процесс формирования (обучения) рабочей модели. В то же время важно отделять индивидуальные особенности распозноваемых объектов от аномальных данных. В настоящей работе предложены логические методы анализа данных, позволяющие провести классификацию данных. В качестве функции классификатора строится функция, которая является логической комбинацией продукионных правил. Она решает ряд проблем, строит все возможные классы, выявляет индивидуальные характеристики объектов, входящих во множество данных, выявляет объекты и их признаки, которые являются выросами. Основываясь на результатах работы построенного классификатора можно выявленные подозрительные объекты дополнительно исследовать на предмет принадлежности множеству выбросов с учетом полученной оценки. Предложенный подход позволяет не только произвести обучающей выборки на классы, но и выявить выбросы, объекты, которые не могут выступать в качестве эталонов обучающей выборки. Предложенный в настоящей работе метод может служить основой для построения процедуры, повышающей информативное качество обучающей выборки в исследуемой предетной области.*

**Ключевые слова:** объект, класс, база знаний, выбросы, информативный вес.

## Введение

В настоящие время существуют хорошо зарекомендовавшие себя методы выявления выбросов. К ним относятся: статистические тесты, модельные тесты, итерационные методы, метрические методы, методы машинного обучения, изолирующие леса и т.д. [1].

В данной работе предлагается логический анализ данных на предмет присутствия в них выбросов и построение робастной процедуры нечувствительной к их влиянию.

**Постановка задачи**

Распознавание образов — это уже целый раздел теоретической информатики, разрабатывающий принципы и методы классификации, а также идентификации предметов, явлений, процессов, сигналов, ситуаций - всех тех объектов, которые могут быть описаны конечным набором некоторых признаков или свойств, характеризующих объект [2].

Описание объекта представляет собой $n$-мерный вектор, где $n$ - число признаков, используемых для характеристики объекта, причем $j$-я координата этого вектора равна значению $j$-го признака, $j = 1,..,n$. В описании объекта допустимо отсутствие информации о значении того или иного признака.

Формальная постановка задачи распознавания образов.

Пусть $X = \{x_1, x_2, ..., x_n\}$ $x_i \in \{0,1,..., k_i - 1\}$, где $k_i \in [2,..., N]$, $N \in \mathbb{N}$ –набор переменных, отражающих признаки характеризующие соответствующий образ. $Y = \{y_1, y_2, ..., y_m\}$ – набор образов, идентифицируемый соответствующими ему признаками. У каждого образа $y_i$ есть соответствующий набор признаков $x_1(y_i),..., x_n(y_i) : y_i = f(x_1(y_i),..., x_n(y_i))$. Иначе говоря $X = \{x_1, x_2, ..., x_n\}$, где $x_i \in \{0,1,..., k_r - 1\}$, $k_r \in [2,..., N], N \in Z$ –входные данные $X_i = \{x_1(y_i), x_2(y_i),..., x_n(y_i)\}$, $i = 1,..., n$, $y_i \in Y, Y = \{y_1, y_2, ..., y_m\}$ – выходные данные:

$$\begin{pmatrix} x_1(y_1) & x_2(y_1) & ... & x_n(y_1) \\ x_1(y_2) & x_2(y_2) & ... & x_n(y_2) \\ ... & ... & ... & ... \\ x_1(y_m) & x_2(y_m) & ... & x_n(y_m) \end{pmatrix} \to \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_m \end{pmatrix}$$

Вид функции $Y = f(X)$ не задан.

Для восстановления этой функции по наблюдениям требуется определить качество данных. Надо заметить, что среди данных могут находиться те, которые могут существенно искажать работу функции классификатора. Как правило, выявления подобных данных проводиться различными статистическими методами, которые хорошо себя зарекомендовали при решении подобных задач и их дальнейшим удалением, либо построением робастных процедур, т.е. процедур нечувствительным к выбросам.

Данные, с которыми приходится иметь дела при решении задач распознавания, как известно, являются не полными, не точными, не

однозначными. Однако получаемые решения должны соответствовать закономерностям, явно и неявно присутствующим в рассматриваемой предметной области. Логические методы могут достаточно хорошо проанализировать данные, выделить существенные и несущественные признаки, выявить минимальный набор правил необходимый для того, чтобы полностью восстановить исходные закономерности. В результате можно получить более компактное и надежное представление исходной информации, обрабатывать которую надежней и быстрее.

Если все правила в рассматриваемой нами области могут быть получены из какой-либо системы, то такая система называется полной.

Класс — это набор образов, объединенных по какой-либо группе признаков.

Каждый объект будет являться представителем одного или нескольких классов, которые могут пересекаться.

Логическую связь между, распознаваемыми, объектами и их признаками, можно представить в как правило продукции следующего вида:

$$\&_{j=1}^{m} x_j(y_i,) \to P(y_i), i = 1,...,l; x_j(y_i) \in \{0,1,...,k-1\},$$

предикат $P(y_i)$ $P(y_i) = 1$ в случае если $y = y_i$ и $P(y_i) = 0$, если $y \neq y_i$ .Это же выражение может быть представлена в следующем виде:

$$\vee_{i=1}^{n} \bar{x}(y_j) \vee P(y_j), j \in [1,\ldots,m]$$

Теперь построим функцию для всех объектов заданной предметной области. Причем функция, которая будет объединять каждое правило, будет конъюнкция, т.е. в логической интерпретации, предметная область — это первое правило и второе правило и…n-е правило одновременно. Тогда функция, которую будем называть решающей имеет вид: $\&_{j=1}^{m} x_j(y_i,) \to P(y_i), i = 1,...,l; x_j(y_i) \in \{0,1,...,k-1\}$

$$f(X) = \&_{j=1}^{m}\left(\vee_{i=1}^{n} \bar{x}_i \vee P(y_j)\right) \quad (1)$$

Можно утверждать в соответствии с функцией (1), что представленная в виде логических соответствий предметная область

$F\left(x_1(y_i),...,x_n(y_i),P^{\sigma}(y_1),...,P^{\sigma}(y_n)\right)$, где $P^{\sigma}(y_i) = \begin{cases} \overline{P(y_i)} & при \quad \sigma = 0 \\ P(y_i) & при \quad \sigma = 1 \end{cases}$,

Будет иметь ложное значение на наборах $\left(x_1(y_i),...,x_n(y_i),P^{\sigma}(y_1),...,\overline{P^{\sigma}(y_i)},...,P^{\sigma}(y_n)\right)$ и истинное на всех остальных на наборах, т.е. это говорит о том, что могут быть любые характеристики у

заявленных образов, и те которые показаны в данных и возможны другие, не может быть только отрицания правила, которое существует в данных.

Построенная таким методом решающая функция, является однозначным интерпретатором исходных данных. На исходных данных функция строит все возможные классы. В соответствии с правилами алгебры логики, после минимизации СДНФ, можно получить наиболее значимые правила для установления закономерностей в исходных данных.

Данная функция обладает рядом свойств и особенностей, она практически строит базу знаний для заданной области данных. Все свойства функции (1) подробно рассмотрены в работах [6].

*Определение.* Логическим описанием класса $K_j$ назовем дизъюнкту, содержащую некоторое множество предикатов, отражающих объекты обучающей выборки и переменные характеризующие признаки этих объектов.

*Утверждение.* Функция

$$f(X) = \&_{i=1}^{n}(\vee_{j=1}^{m} \overline{x}(y_j) \vee y_j), \qquad x(y_j) \in [0,1,2], \quad y_j \in Y$$

полна на заданном множестве признаков.

**Алгоритм моделирования объектной части решающей функции**

Для реализации объектной части решающей функции можно воспользоваться следующим алгоритмом:

- рассмотрим Таблицу, столбцами которой будут являться признаки, каждый со своей значностью

- строками, будут являться объекты, расписанные в соответствии со своими характеристиками. Например, если объект $y_1$ по первой переменной имеет значение «1» то помещаем его в столбец $1_1$ (Таблица 3).

Таблица 1.

| $0_1$ | $1_1$ | $k_1-1$ | $0_2$ | $1_2$ | $k_2-1$ | … | $0_n$ | $1_n$ | $k_n-1$ |
|---|---|---|---|---|---|---|---|---|---|
| | $y_1$ | | | | | … | | | $y_1$ |
| | $y_2$ | | | | | … | | $y_2$ | |
| … | … | … | … | … | … | … | … | … | … |
| $y_m$ | | | | $y_m$ | | … | | | $y_m$ |

- смотрим каждый столбец Таблицы. Если в столбце более одного элемента, то по данному признаку объекты образуют класс и можно, в следующей свободной строке выписать этот класс. (Таблица 2).

Таблица 2.

| $0_1$ | $1_1$ | $k_1-1$ | $0_2$ | $1_2$ | $k_2-1$ | … | $0_n$ | $1_n$ | $k_n-1$ |
|---|---|---|---|---|---|---|---|---|---|
| | ~~$y_1$~~ | | | | | … | | | ~~$y_1$~~ |
| | ~~$y_2$~~ | | | | | … | | $y_2$ | |
| | $y_1y_2$ | | | | | | | | |
| … | … | … | … | … | … | … | … | … | … |
| $y_m$ | | | | $y_m$ | | … | | | ~~$y_m$~~ |
| | | | | | | | | | $y_1y_m$ |

- если остались строки с одиночными, невычеркнуыми объектами, то объект идентифицируется именно по этим переменным, это его индивидуальные признаки. Набор таких индивидуальных признаков, является наиболее существенными правила для заданных данных.

Строки, у которых несколько объектов в одном столбце, демонстрируют классы, которые возможно получить исследуя данную предметную область.

**Пример.** Пусть заданный набор данных характеризуется следующей Таблицей:

Таблица 3

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| *0* | *1* | *2* | *0* | *a* |
| *0* | *2* | *1* | | *в* |
| *1* | *0* | *1* | *1* | *c* |
| *1* | *1* | *1* | *1* | *d* |
| *0* | *0* | *0* | *2* | *e* |

Не строя всю функцию классификатор целиком. Построим для наглядности только ее объектную часть в соответствии с выше описанным алгоритмом (Таблица 4).

Таблица 4

| $x_1$ | | $x_2$ | | | $x_3$ | | | $x_4$ | | | $K$ (классы) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | |
| ~~a~~ | | | ~~a~~ | | | | a | ~~a~~ | | | |
| ~~b~~ | | | | b | | ~~b~~ | | ~~b~~ | | | |
| ~~*ab*~~ | | | | | | | | *ab* | | | $ab\ x_1^0\ x_4^0$ |
| | ~~c~~ | ~~c~~ | | | | ~~c~~ | | | ~~c~~ | | |
| | | | | | | ~~bc~~ | | | | | ~~$bcx_3^1$~~ |
| | ~~d~~ | | ~~d~~ | | | ~~d~~ | | | ~~d~~ | | |
| | | | | | | bcd | | | | | $bcdx_3^1$ |
| | cd | | | | | | | | cd | | $cd\ x_1^1x_4^1\ x_3^1$ |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ad | | | | | | | | $ad\ x_2^1$ |
| ~~e~~ | | ~~e~~ | | | e | | | | | e | |
| abe | | | | | | | | | | | $abe\ x_1^0$ |
| | | ce | | | | | | | | | $ce\ x_2^0$ |

После работы алгоритма мы получили все возможные классы на рассматриваемых данных: $\{abx_1^0x_4^0,\ bcdx_3^1,\ cdx_1^1x_4^1x_3^1,\ adx_2^1,\ abex_1^0,cex_2^0,\ bx_2^2,ax_3^2,\ ex_3^0x_4^2\}$

Объектная часть, описанной ранее функции будет выглядеть следующим образом

$$f_2(X) = abx_1^0x_4^0 \vee bcdx_3^1 \vee cdx_1^1x_4^1x_3^1 \vee adx_2^1 \vee abex_1^0 \vee cex_2^0 \vee bx_2^2 \vee ax_3^2 \vee ex_3^0x_4^2$$

*Определение.* Число объектов объеденных в класс по совокупности признаков назовем объектным весом класса $(v_{об})$.

$$v_{об}(bcdx_3^1)=3.$$

*Определение.* Число признаков, объединяющих в класс определенное количество объектов, назовем признаковым весом $(v_{приз})$.

$$v_{приз}(cdx_1^1x_4^1x_3^1)=3$$

В рамках данной работы, претендентами на выбросы будем называть объекты, которые не входят в основные классы.

Для данного примера это $bx_2^2; ax_3^2; ex_3^0x_4^2$.

Заметим, что причиной появления классов, включающих в себя малое количество объектов, могут быть следующие:

Это могут объекты, характеризующие новое знание или это, могут быть искаженные по тем или иным причинам данные, т.е. - выбросы.

Введем понятие информативности класса, представленного конъюнкциями. Класс – это множество объектов и характеризующих их признаков. Для информативной характеристики совокупности количества объектов класса и количества, определяемых этот класс признаков введем следующую величину информативный вес: $w_i = \frac{n^i}{n} + \frac{m^i}{m}$ где $n^i$ - число признаков, определяющих класс $K^i$, $n$- общее число признаков, $m_i$ - число объектов попадающих в класс $K^i$, $m$- общее число объектов исследуемой предметной области.

Для разбиения предметной области на классы, а также определения в ней выбросов относительно полученного разбиения предлагается следующий алгоритм:

1. Определить информационный вес каждой конъюнкции, полученной в результате работы предложенного выше алгоритма.
2. Выбрать конъюнкции $K^i$ с наибольшим весом $w^i$.
3. Вычеркнуть все конъюнкции, содержащие объекты, входящие в класс $K^i$
4. Перейти к шагу 2.
5. После того как остались непересекающиеся по объектам классы, находим число классов $L = \mathrm{mod} \sum_i w_i$.

6. Если существует $K^j$ такой, что $w_j$ не влияет на $L$, то $K^j$ является выбросом для данной классификации.

Для данного примера $n=4$ (четыре признака), $m=5$(пять объектов)
Шаг 1. Считаем информативные веса у каждой конъюнкции, находим наибольший вес и вычеркиваем все конъюнкции, содержащие элементы конъюнкции с наибольшим весом:

| К (класс, конъюнкция) | $w$ |
|---|---|
| $abx_1^0x_4^0$ | $\frac{2}{5}+\frac{2}{4}=\frac{18}{20}$ |
| ~~$bcd\ x_3^1$~~, | $\frac{17}{20}$ (содержит cd) |
| $cdx_1^1x_4^1x_3^1$ | $\frac{23}{20}$ - максимальноезначение веса |
| ~~$adx_2^1$~~ | $\frac{13}{20}$ ( содержит d) |
| $abex_1^0$ | $\frac{17}{20}$ |
| ~~$cex_2^0$~~ | $\frac{13}{20}$ ( содержит c) |
| | |
| $bx_2^2$ | $\frac{9}{20}$ |
| $ax_3^2$ | $\frac{9}{20}$ |
| $ex_3^0x_4^2$ | $\frac{14}{20}$ |

На первом шаге определен класс $cdx_1^1x_4^1x_3^1$с весом $w^{cd}=\frac{23}{20}$.

Шаг 2. Из оставшихся конъюнкций находим с наибольшим весом и проводим те же операции:

| К (класс, конъюнкция) | *w* |
|---|---|
| $abx_1^0x_4^0$ | $\frac{2}{5}+\frac{2}{4}=\frac{18}{20}$ максимальное значение веса |
| ~~$abex_1^0$~~ | $\frac{17}{20}$ (содержит ab) |
| ~~$bx_2^2$~~ | $\frac{9}{20}$ (содержит b) |
| ~~$ax_3^2$~~ | $\frac{9}{20}$ (содержит a) |
| $ex_3^0x_4^2$ | $\frac{14}{20}$ |

На втором шаге получили класс $abx_1^0x_4^0$ с весом $w^{ab}=\frac{18}{20}$ и $ex_3^0x_4^2$ с $w^e=\frac{14}{20}$.

$$L=\mathrm{mod}(\mathrm{w}^{\mathrm{cd}}+w^{ab}+w^e)=\mathrm{mod}(\frac{23}{20}+\frac{18}{20}+\frac{14}{20})=2$$

В данном примере для введенной нами весовой информативности мы получаем будет два класса. Это будут классы с наибольшими весами: $cdx_1^1x_4^1x_3^1$ и $abx_1^0x_4^0$.

Причем $L=\mathrm{mod}(\mathrm{w}^{\mathrm{cd}}+w^{ab})=\mathrm{mod}(\frac{23}{20}+\frac{18}{20})=2$.

И так же выброс $ex_3^0x_4^2$ для полученной классификации. Как видно информативный вес объекта *c*не влияет на количество классов, поэтому его можно трактовать как выброс.

Поэтому можно говорить, что представленные данные поддаются следующее классификации: $cdx_1^1x_4^1x_3^1$и $abx_1^0x_4^0$.

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | *y* |
|---|---|---|---|---|
| ***0*** | *1* | *2* | ***0*** | ***a*** |
| ***0*** | *2* | *1* | ***0*** | ***в*** |
| *1* | *0* | *1* | *1* | *c* |
| *1* | *1* | *1* | *1* | *d* |
| *0* | *0* | *0* | *2* | *e* |

А объект *e*, для выбранной классификации будет являться выбросом.

***Утверждение.*** Объект $y_i \in Y, Y = \{y_1, y_2, ..., y_m\}$ - множество классифицируемых объектов будет являться выбросом в рамках модели, в которой величина информативного веса представлена как

$$w_i = \frac{n^i}{n} + \frac{m^i}{m}$$

где $n^i$ - число признаков, определяющих класс $K^i$,

$n$- общее число признаков,

$m_i$ - число объектов попадающих в класс $K^i$,

$m$- общее число объектов исследуемой предметной области, если

$$L = \operatorname{mod} \sum_i w_i = \operatorname{mod}(\sum_i w_i - w^{y_i}),$$

где $i$- номер класса, полученный после логической классификации заданной предметной области.

**Заключение**

Анализ исходных данных – важный процесс для построения модели зависимостей в заданной предметной области. Качество модели может сильно пострадать в случае, если объекты с искаженной информацией рассматриваются, как объекты, отражающие действительную зависимость, которую стараются найти. Поэтому необходимо иметь приоритеты, по которым будет проходить построение моделей. Для задач классификации, в случае, когда каждый признак равнозначный по своей важности для идентификации класса, можно воспользоваться предложенным методом, который позволяет выявить число классов для заданной предметной области, а также найти объекты, не принадлежащие этим классам, если таковые есть.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №18-01-00050-а.

## ЛИТЕРАТУРА

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. Журавлёв Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации//Проблемы кибернетики. 1978. Т. 33. С. 5–68.

3. Лютикова Л.А., Шматова Е.В. Анализ и синтез алгоритмов распознавания образов с использованием переменно-значной логики // Информационные технологии. №4. Том 22. 2016. С. 292—297.
4. Лютикова Л.А., Шматова Е.В. Логический подход к коррекции результатов работы $\Sigma\Pi$-нейронных сетей//Информационные технологии. 2018. Т. 24. №2. С. 110-116.
5. Шибзухов З.М. О принципе минимизации эмпирического риска на основе усредняющих агрегирующих функций//Доклады РАН. 2017. Т.476. №5. С. 495-499.
6. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: МДК Пресс, 2015. 400 с.

L.A. Lyutikova

## CONSTRUCTION OF A LOGICAL ALGORITHM FOR DETECTING EMISSIONS INTO A DISTURBABLE DATA

*Institute of Applied Mathematics and Automation of Kabardino-Balkarian Scientific Center, RAS, Russia*
*(IAMA KBSC RAS)*

*The paper proposes a logical approach to data quality analysis for solving machine-learning problems. When developing machine-learning algorithms, a part of the initial data of the problem being solved is combined into a training sample. As a rule, the quality of this data is not ideal, and this is a rather acute problem arising in the construction of training recognition systems. Since the construction of the recognition model is the result of the sequential presentation of the initial data set, their incorrectness can significantly distort the final model, which stresses the results of the recognition algorithms. The data that introduce distortions in building a model is called outliers. The cause of emissions is the interference of the equipment, incorrect interpretation of the expert, noise, etc. In this regard, the task of analyzing data to identify emissions and reducing their influence on the process of formation (training) of the working model arises. At the same time, it is important to separate the individual features of recognized objects from abnormal data. In the present work, logical methods of data analysis are proposed, allowing data to be classified. As a classifier function, a function is constructed that is a logical combination of production rules. It solves a number of problems, builds all possible classes, reveals the individual characteristics of objects included in the data set, identifies objects and their signs that are grown. Based on the results of the constructed classifier, the identified suspicious objects can be additionally investigated for belonging to a set of emissions, taking into account the obtained estimate. The proposed approach allows not only to make a training sample for classes, but also to identify emissions, objects that can not act as standards of the training sample. The method proposed in this paper can serve as the basis for constructing a procedure that enhances the informative quality of a training sample in the pre-project area under study.*

**Keywords**: object, class, knowledge base, emissions, informative weight.

## REFERENCES

1. D'yakonov A.G., Golovina A.M. Vyyavlenie anomaliy v rabote mekhanizmov metodami mashinnogo obucheniya // Analitika i upravlenie dannymi v oblastyakh s intensivnym ispol'zovaniem dannykh: trudy XIX Mezhdunarodnoy konferentsii DAMDID/RCDL'2017, 2017. pp. 469–476.
2. Zhuravlev Yu. I. Ob algebraicheskom podkhode k resheniyu zadach raspoznavaniya ili klassifikatsii // Problemy kibernetiki. 1978. Vol. 33. pp. 5–68.
3. Lyutikova L. A., Shmatova E. V. Analiz i sintez algoritmov raspoznavaniya obrazov s ispol'zovaniem peremenno-znachnoy logiki // Informatsionnye tekhnologii. No.4. Vol. 22. 2016. pp. 292—297.
4. Lyutikova L.A., Shmatova E.V. Logicheskiy podkhod k korrektsii rezul'tatov raboty $\Sigma\Pi$-neyronnykh setey // Informatsionnye tekhnologii. 2018. Vol. 24. No.2. pp. 110-116.
5. Shibzukhov Z.M. O printsipe minimizatsii empiricheskogo riska na osnove usrednyayushchikh agregiruyushchikh funktsiy. // Doklady RAN. 2017. Vol.476. No.5. pp. 495-499.
6. Flakh P. Mashinnoe obuchenie. Naukai isskustvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannykh. M.: MDK Press, 2015. 400 p.