ИНФОРМАТИКА. ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА. УПРАВЛЕНИЕ

УДК 004.8; 519.7
DOI: 10.35330/1991-6639-2019-6-92-67-74

МЕТОД ОБНАРУЖЕНИЯ ВЫБРОСОВ В ДАННЫХ НА ОСНОВЕ МНОГОЗНАЧНОЙ ЛОГИКИ ПРЕДИКАТОВ*

Л.А. ЛЮТИКОВА[1,2], М.А. ШОГЕНОВ[1]

[1] Институт прикладной математики и автоматизации –
филиал ФГБНУ «Федеральный научный центр
«Кабардино-Балкарский научный центр Российской академии наук»
360000, КБР, г. Нальчик, ул. Шортанова, 89 А
E-mail: ipma@niipma.ru
[2] ФГБНУ «Федеральный научный центр
«Кабардино-Балкарский научный центр Российской академии наук»
360002, КБР, г. Нальчик, ул. Балкарова, 2
E-mail: kbncran@mail.ru

В работе рассматривается метод анализа данных, которые будут использованы при решении задач машинного обучения, на предмет нахождения в этих данных шумов и неточностей, искажений, которые препятствуют построению адекватной модели. Данные такого рода называются выбросами. Предлагаемый подход использует методы и алгоритмы, основанные на системах многозначной логики. Многозначную логику можно использовать в случае с многомерными разнородными признаками, характеризующими объекты исходной предметной области. Для проведения качественного анализа данных в работе предлагается следующий порядок действий: строится многозначная логическая функция для анализируемых данных, которая находит все возможные классы на рассматриваемой предметной области; далее проводится анализ объектов, которые не попали в построенные классы по ряду признаков; проверяется гипотеза о том, что данные объекты являются выбросами. В работе проверка гипотезы – это последовательность логических правил для восстановления исходных зависимостей, представленных в обучающей выборке. Предлагаемый подход рассматривался для задач классификации в случае многомерных дискретных признаков, когда каждый признак может принимать k различных значений и являться равнозначным по своей важности для идентификации класса.

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

Введение

Качество данных – одна из важнейших проблем, возникающих при решении интеллектуальных задач. Наличие в обучающей выборке искаженных данных влияет на конечную работу алгоритма, поскольку такие данные искажают модель, построенную в результате обработки исходной предметной области, что в свою очередь способно сильно повлиять на правильность принимаемого решения. Причина появления таких данных зависит от исследуемой области, это может быть сбой аппаратуры в случае распознавания информации с датчиков, это могут быть ошибки эксперта в случае моделирования рассуждения эксперта и другие причины.

Выбросы – резко отличающиеся признаки объектов или наблюдения в наборе данных. Вообще при анализе данных шумы и выбросы являются достаточно общей проблемой. Поэтому необходимо их обнаружить и оценить степень влияния на результаты дальнейшего анализа.

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

В настоящие время разработаны различные методы борьбы с выбросами: статистические тесты, модельные тесты, итерационные методы, метрические методы, методы машинного обучения, изолирующие леса и т.д. [1-4].

В данной работе предлагается метод выявления выбросов, основанный на построении логического классификатора.

Постановка задачи

Классифицируемый, или распознаваемый, объект будет представлен n-мерным вектором, n – число характерных признаков рассматриваемого объекта, j-я координата этого вектора равна значению j-й характеристики, $j=1,..,n$. Информация о какой-либо характеристике объекта может отсутствовать. Мерность рассматриваемого свойства объекта $k_i \in [2,...,N], N_i$ зависит от метода кодирования *i*-й характеристики [9].

Постановка задачи:

Пусть $X=\{x_1,x_2,...,x_n\}$ $x_i \in \{0,1,...,k_i-1\}$, где $k_i \in [2,...,N]$, $N \in Z$ – набор свойств, которые характеризуют заданный объект. $Y=\{y_1,y_2,...,y_m\}$ – множество рассматриваемых объектов. Для каждого объекта y_i есть соответствующий набор признаков $x_1(y_i),...,x_n(y_i): y_i = f(x_1(y_i),...,x_n(y_i))$. Или $X=\{x_1,x_2,...,x_n\}$, 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:

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

Необходимо построить функцию такую, что $Y=f(X)$.

Понятно, что среди данных могут находиться те, которые будут существенно искажать построение классификатора при различных подходах. Поэтому встает вопрос об их обнаружении, дальнейшем удалении или построении робастных процедур, т.е. процедур, нечувствительных к выбросам. Использование математической логики удобно для построения такого рода процедур [7,8].

Рассмотрим необходимые определения:

Определение 1. Функция обеспечивает свойство полноты для получаемых решений, если она обеспечивает вывод всех возможных решений.

Определение 2. Получаемые решения называются непротиворечивыми, если на наборе признаков $X_j \in X$, где X – пространство признаков, невозможно получить выводы: $X_j \rightarrow y_j$ и $X_j \rightarrow \sim y_j$.

Определение 3. Класс – множество объектов, обладающих определенным свойством или набором свойств [5,6].

Поскольку каждый рассматриваемый объект характеризуется рядом признаков, а каждый признак имеет несколько значений, которые отражаются в виде значений переменных $x_i \in \{0,1,...,k_i-1\}$, введем понятие инверсии для многозначных систем. В нашем случае отрицанием какого-либо значения будет являться любое другое значение из числа заданных, кроме отрицаемого.

$$\overline{x^j} = x^0 \vee x^1 \vee \ldots \vee x^{j-1} \vee x^{j+1} \vee \ldots \vee x^{k-1}, \text{ где } x^j=j.$$

Основные операции:

$$0\,\&\,X=0,\ 1\&\,X=X,\ \ (k-1)X=(k-1),\ 0X=X,\ \ x^j\,\&\,x^k=\begin{cases}x^j, & j=k\\ 0, & j\neq k\end{cases}$$

Состояние, когда некоторый объект характеризуется заданным набором признаков, каждый из которых имеет определенное значение, представимо следующим правилом:

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

$P(y_i)$ – предикат, который $P(y_i)=1$, если $y=y_i$ и $P(y_i)=0$, если $y\neq y_i$.

Определение 4. Решающим правилом при построении искомой функции назовем следующее выражение:

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

где предикат $P(y_i)$ принимает значение «истина», т.е. $P(y_i)$=1 в случае, если $y=y_i$ и $P(y_i)$=0, если $y\neq y_i$.

Перепишем это правило через функции дизъюнкции, конъюнкции и отрицания, получим правило, определяющие объект y_j по его признакам $x_i(y_j)$:

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

Тогда функция, описывающая совокупность всех заданных объектов и их признаков, будет следующая:

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

Обрабатываемые данные могут быть представлены булевой функцией от $m+n$ переменных:

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

Такая функция будет принимать значение «ложь» на наборах $\left(x_1(y_i),...,x_n(y_i),P^{\sigma}(y_1),...,\overline{P^{\sigma}(y_i)},...,P^{\sigma}(y_n)\right)$ там, где есть признаки объекта y_j, но отрицается сам объект, и «истина» в остальных случаях.

Построенный классификатор является логической функцией и может быть минимизирован по правилам минимизации ДНФ.

Построенная функция обладает рядом свойств [12].

Определение. Логическим описанием класса K_j назовем дизъюнкту, состоящую из конечного числа предикатов, отражающих объекты обучающей выборки, и переменных, характеризующих признаки этих объектов.

Утверждение. Функция

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

полна на заданном множестве данных.

Структура функции представляет собой дизъюнкты, содержащие предикаты, т.е. объекты, набор таких дизъюнктов назовем объектной частью функции, и дизъюнкт, состоящий из одних переменных, т.е. признаков., эту часть функции назовем признаковой.

По объектной части функции можно восстановить всю обучающую выборку, по признаковой – определить, совокупность каких признаков не была задействована в описании рассматриваемого множества объектов [11].

Алгоритм построения объектной части логического классификатора

Для построения объектной части рассматриваемой выше функции предлагается следующий алгоритм:

- создадим таблицу, столбцами таблицы будут являться значности соответствующих признаков;

- строками в данной таблице являются объекты, разносятся по соответствующим значениям. Например, если объект y_1 по первой переменной имеет значение «1», то помещаем его в столбец 1_1 (табл. 1);

Таблица 1

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

- далее рассматриваем столбцы таблицы. Если в столбце более одного элемента, то по данному признаку объекты образуют класс, и можно в следующей свободной строке выписать этот класс. (табл. 2);

Таблица 2

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

- если остались строки с одиночными, невычеркнутыми объектами, то объект идентифицируется именно по этим переменным, это его индивидуальные признаки. Набор таких индивидуальных признаков является наиболее существенным правилом 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 3:

Таблица 3

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*

Не строя всю функцию-классификатор целиком, построим для наглядности только ее объектную часть в соответствии с вышеописанным алгоритмом (табл. 4).

Таблица 4

x_1		x_2			x_3			x_4			*К (классы)*
0	1	0	1	2	0	1	2	0	1	2	
~~a~~			~~a~~				a	~~a~~			$a\,x_3^2$
~~b~~				b		~~b~~		~~b~~			$b\,x_2^2$
~~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, объедененных 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 *($v_{приз}$):*

$$V_{приз}(cd\ x_1^1x_4^1x_3^1)=3.$$

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 причинами появления классов, включающих в себя малое количество объектов, могут быть следующие.

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

В нашем примере $x_1 \in [0,1] x_2 \in [0,1,2] x_3 \in [0,1,2] x_4 \in [0,1,2]$.

Выпишем классы в порядке возрастания объектных весов, восстанавливая при этом исходные данные (табл. 5):

Таблица 5

$abx_1^0x_4^0$				
x_1	x_2	x_3	x_4	y
0			0	a
0			0	в
				c
				d
				e

$bcd\ x_3^1$				
x_1	x_2	x_3	x_4	y
0			0	a
0		1	0	в
		1		c
		1		d
				e

$cdx_1^1x_4^1x_3^1$				
x_1	x_2	x_3	x_4	y
0			0	a
0		1	0	в
1		1	1	c
1		1	1	d
				e

adx_2^1				
x_1	x_2	x_3	x_4	y
0	1		0	a
0		1	0	в
1		1	1	c
1	1	1	1	d
				e

$abex_1^0$				
x_1	x_2	x_3	x_4	y
0	1		0	a
0		1	0	в
1		1	1	c
1	1	1	1	d
0				e

cex_2^0				
x_1	x_2	x_3	x_4	y
0	1		0	a
0		1	0	в
1	0	1	1	c
1	1	1	1	d
0	0			e

Остались одиночные элементы $bx_2^2; ax_3^2; ex_3^0x_4^2$. Поскольку элемент b не входит ни в один из классов по переменной x_2, это значит, что x_2 в случае элемента b не принимает значения *«1» и «0»,* можно однозначно утверждать, что $x_2 = 2$ для элемента b, т.к. $x_2 \in [0,1,2]$.

То же можно утверждать для объекта e в случае переменной x_4.

Достроенная таблица имеет вид (табл. 6).

Таблица 6

x_1	x_2	x_3	x_4	Y
0	1		0	A
0	2	1	0	B
1	0	1	1	C
1	1	1	1	D
0	0		2	E

Однако заполнить пустые ячейка для a и e мы не сможем, если не будем знать значение переменной x_3 для одного из элементов. Следовательно, эти объекты не попадают под общую логику классификации рассматриваемых данных и являются выбросами.

Логический анализ заданной предметной области на предмет обнаружения в ней выбросов может выглядеть следующим образом.

1. Строим логическую функцию-классификатор или используем алгоритм для построения объектной части логической функции для выявления всех возможных классов в заданной предметной области.

2. Из построенного набора всех возможных классов убираем классы с минимальным объектным весом.

3. По оставшемуся набору восстанавливаем исходные данные.

4. Если данные не удается восстановить полностью, то элементы в удаленных классах являются выбросами.

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 многомерных дискретных признаков, 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.

ЛИТЕРАТУРА

1. Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka GrabskaBarwinska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. Nature, 2016. № 538(7626). Pp. 471-476.

2. Ashley I. Naimi, Laura B. Balzer Stacked generalization: an introduction to super learning // European Journal of Epidemiology (2018) 33:459-464.

3. Fan Yang Zhilin Yang William W. Cohen Differentiable Learning of Logical Rules for Knowledge Base Reasoning // Advances in Neural Information Processing Systems. Volume 2017-December, 2017, Pages 2320-2329.

4. Peter Flach Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, 2012. 396 p.

5. Rahman Akhlaqur & Tasnim Sumaira. Ensemble Classifiers and Their Applications: A Review // International Journal of Computer Trends and Technology. (2014). Vol. 10. No 1. Pp. 31-35.

6. Serafini L., Garcez A.A. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. Preprint arXiv:1606.04422, 2016 - arxiv.org

7. Shibzukhov Z.M. Correct Aggregation Operations with Algorithms // Pattern Recognition and Image Analysis. 2014, Vol. 24, No. 3. Pp. 377-382.

8. Shibzukhov, Z.M. On the principle of empirical risk minimization based on averaging aggregation functions. Doklady Mathematics, 2017. Volume 96, Issue 2. Pp. 494-497.

9. Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, Eric Xing. Harnessing Deep Neural Networks with Logic Rules // Computer Science Learning. 2016. P. 2410-2420. arXiv:1603.06318.

10. *Дюкова Е.В., Журавлев Ю.И., Прокофьев П.А.* Методы повышения эффективности логических корректоров // Машинное обучение и анализ данных. 2015. Т. 1. № 11. С. 1555-1583.

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

12. *Лютикова Л.А., Шматова Е.В.* Анализ и синтез алгоритмов распознавания образов с использованием переменнозначной логики // Информационные технологии. Том 22. № 4. 2016. С. 292-297.

REFERENCES

1. Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka GrabskaBarwinska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. Nature, 2016. № 538(7626). Pp. 471-476.

2. Ashley I. Naimi, Laura B. Balzer Stacked generalization: an introduction to super learning // European Journal of Epidemiology (2018) 33:459-464.

3. Fan Yang Zhilin Yang William W. Cohen Differentiable Learning of Logical Rules for Knowledge Base Reasoning // Advances in Neural Information Processing Systems. Volume 2017-December, 2017, Pp. 2320-2329.

4. Peter Flach Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, 2012. 396 p.

5. Rahman Akhlaqur & Tasnim Sumaira. Ensemble Classifiers and Their Applications: A Review // International Journal of Computer Trends and Technology. (2014). Vol. 10. No 1. Pp. 31-35.

6. Serafini L., Garcez A.A. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. Preprint arXiv:1606.04422, 2016 - arxiv.org

7. Shibzukhov Z.M. Correct Aggregation Operations with Algorithms // Pattern Recognition and Image Analysis. 2014, Vol. 24, No. 3. Pp. 377-382.

8. Shibzukhov, Z.M. On the principle of empirical risk minimization based on averaging aggregation functions. Doklady Mathematics, 2017. Volume 96, Issue 2. Pages 494-497.

9. Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, Eric Xing. Harnessing Deep Neural Networks with Logic Rules // Computer Science Learning. 2016. P. 2410-2420. arxiv:1603.06318.

10. Dyukova E.V., Zhuravlev Yu.I., Prokof`ev P.A. *Metody` povy`sheniya e`ffektivnosti logicheskix korrektorov* [Methods for increasing the efficiency of logical correctors] // *Mashinnoe obuchenie i analiz danny`x* [Machine Learning and Data Analysis]. 2015. T. 1. № 11. Pp. 1555-1583.

11. Zhuravlev Yu.I. *Ob algebraicheskom podxode k resheniyu zadach raspoznavaniya ili klassifikacii* [On the algebraic approach to solving recognition or classification problems] // *Problemy` kibernetiki* [Problems of Cybernetics].1978. T. 33. Pp. 5-68.

12. Lyutikova L.A., Shmatova E.V. *Analiz i sintez algoritmov raspoznavaniya obrazov s ispol`zovaniem peremenno-znachnoj logiki* [Analysis and synthesis of pattern recognition algorithms using variable-valued logic] // *Informacionny`e texnologii* ["Information Technologies"]. Tom 22. № 4. 2016. Pp. 292-297.

OUTLIERS DETECTION METHOD FOR DATA BASED ON MULTI-VALUED PREDICATE LOGIC

L.A. LYUTIKOVA[1,2], M.A. SHOGENOV[1]

[1] Institute of Applied Mathematics and Automation –
branch of the FSBSE "Federal Scientific Center
"Kabardin-Balkar Scientific Center of the Russian Academy of Sciences"
360000, KBR, Nalchik, Shortanov street, 89 "A"
E-mail: ipma@niipma.ru
[2] FSBSI " Federal scientific center
"Kabardin-Balkar scientific center of the Russian academy of sciences"
360002, KBR, Nalchik, 2 Balkarova st.
E-mail: lab.msb@mail.ru

The paper discusses a method for analyzing data that will be used in solving machine learning problems to find noise and inaccuracies, distortions in these data that impede the construction of an adequate model. Data of this kind is called outliers. The proposed approach uses methods and algorithms based on multi-valued logic systems. Multivalued logic can be used in the case of multidimensional heterogeneous features that characterize the objects of the original subject area. To conduct a qualitative data analysis, the following procedure is proposed in the work: a multi-valued logical function is constructed for the analyzed data, which finds all possible classes on the subject area under consideration; Further, the analysis of objects that did not fall into the constructed classes for a number of signs; and the hypothesis that these objects are emissions is tested. In the work, a hypothesis test is a sequence of logical rules for restoring the original dependencies presented in the training set. The proposed approach was considered for classification problems, in the case of multidimensional discrete features, where each feature can take k-different values and be equivalent in importance to class identification.

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

Работа поступила 08.11.2019 г.