УДК 519.688

УСТОЙЧИВОСТЬ РАЗБИЕНИЯ ДАННЫХ НА ИНТЕРВАЛЫ В ЗАДАЧАХ РАСПОЗНАВАНИЯ И ПОИСК СКРЫТЫХ ЗАКОНОМЕРНОСТЕЙ

© 2018 Е.Н. Згуральская

Институт авиационных технологий и управления
Ульяновского государственного технического университета

Статья поступила в редакцию 01.11.2018

Большую роль для совершенствования цифровых технологий в научной, производственной и социальной сферах имеет поиск новых знаний, содержащихся в базах и хранилищах данных в форме скрытых закономерностей. В данной работе для выявления скрытых закономерностей при распознавании объектов рассматривается метод разбиения значений признаков на непересекающиеся интервалы. В качестве критерия качества предлагается использовать значение показателя устойчивости разбиения исходных и латентных признаков на интервалы.
*Ключевые слова: С*крытые закономерности, устойчивость разбиения на интервалы, интеллектуальный анализ данных.

ВВЕДЕНИЕ

Использование цифровых технологий в научной, производственной (в частности, в авиации) и социальной сферах являются одним из главных факторов инновационного развития современного общества. Важную роль для совершенствования цифровых технологий играют информационные модели, основанные на знаниях. Как правило, неявные знания содержатся в базах и хранилищах данных в форме скрытых закономерностей. Поиск скрытых закономерностей является основной целью разработки и реализации методов интеллектуального анализа данных (ИАД).

К числу основных проблем построения информационных моделей в слабо структурированных предметных областях относятся выбор описаний допустимых объектов и высокая комбинаторная сложность алгоритмов для поиска логических закономерностей. На решение этих проблем ориентирована разработка методов поиска информативных наборов признаков и подмножества объектов обучения, которые обладают лучшим качеством в смысле решения задач распознавания, чем исходные множества признаков и объектов [1]. На базе методов ИАД разрабатываются информационные модели для объяснения процесса интуитивного принятия решений.

Проблемы имеются в выборе способов предобработки данных с целью уменьшения комбинаторной сложности алгоритмов ИАД, в разработке способов повышения обобщающей способности алгоритмов распознавания, связанных с выбором оптимальных по мощности наборов признаков в описании допустимых объектов. Отсутствие ограничений на число признаков в наборах может привести к явлению, которое Беллман назвал «проклятие размерности» [2].

Существует потребность в разработке и обосновании новых эвристик и критериев для проверки истинности гипотезы о компактности классов [1] при распознавании образов в рамках информационных моделей, в использовании новых методов визуализации для анализа отношений между объектами. Для удовлетворения такой потребности предлагается использовать интервальные методы анализа данных [3]. Границы интервалов определяются как для исходных и латентных признаков, так и для значений мер близости между объектами и признаками.

Одним из универсальных ограничений на использование интервальных методов является инвариантность к масштабам измерений данных. Важность свойства инвариантности выражается в однозначности интерпретации результатов алгоритмов ИАД в рамках информационной модели предметной области. Свойство инвариантности даёт возможность для:

- выбора латентных признаков при моделировании процесса интуитивного принятия решений;
- визуализации описаний объектов из разнотипного признакового пространства;
- упорядочивания разнотипных признаков по отношению информативности.

В статье рассматривается интервальный метод анализа данных, применяемый для задач

Згуральская Екатерина Николаевна, старший преподаватель кафедры «Самолетостроение».
E-mail: iatu@inbox.ru

распознавания с непересекающимися классами. Целью анализа является обнаружение скрытых закономерностей в данных, которые легко представить как новое знание в наглядной для пользователя форме. Новизна знаний выражается в том, что они не являются подтверждением раннее полученных сведений.

1. ОПИСАНИЕ МЕТОДА РАЗБИЕНИЯ ЗНАЧЕНИЙ ПРИЗНАКОВ НА ИНТЕРВАЛЫ И ОЦЕНКИ КАЧЕСТВА РАЗБИЕНИЯ

Предлагается метод определения непересекающихся интервалов количественных признаков, в границах которых доминируют значения объектов одного из непересекающихся классов. На базе этого метода стало возможным как вычисление обобщённых оценок объектов (латентных признаков) в разнотипном признаковом пространстве, так и меры их устойчивости.

Пусть дано множество M допустимых объектов, разбитое на l непересекающихся подмножеств (классов) $K_1,\dots,K_l$. Считается, что представители классов заданы через выборку (подмножество M) объектов $E_0 = \{S_1, \dots, S_m\}$. Объекты выборки описываются с помощью n разнотипных признаков, из которых ξ измеряются в интервальных шкалах, а $n - \xi$ в номинальных.

Вычисление устойчивости объектов по значениям исходных и латентных признаков производится относительно отдельных классов. Необходимость сведения решения к двухклассовой задачи распознавания с объектами из K_t и $CK_t = M \setminus K_t$, $t=1,\dots,l$ связана с тем, что:

-значение любого количественного признака (исходного и латентного) относительно. Объекты каждого из классов противопоставляются объектам противоположных классов (например, класс заболевших и умерших от сердечно-сосудистых заболеваний противопоставляется классу практически здоровых людей);

- отсутствуют наборы аналитических функций для восстановления зависимостей в пространстве разнотипных признаков.

Требуется:

- на множестве допустимых значений каждого из количественных признаков определить разбиение на минимальное число непересекающихся интервалов, в границах которых доминируют значения объектов класса K_t или $CK_t = M \setminus K_t$, $t=1,\dots,l$;

-вычислить значения меры устойчивости разбиения на интервалы признаков объектов E_0 относительно класса K_t, $t=1,\dots,l$.

Обозначим через I,J множество номеров соответственно количественных и номинальных (качественных) признаков $X = \{x_1,\dots,x_n\}$ в описании допустимых объектов, $|I| + |J| = n$. Для удобства выкладок будем рассматривать два класса объектов K_1 и K_2.

Произведём выбор интервалов для каждого количественного признака, в границах которых доминируют значения объектов класса K_t или K_{3-t}, $t=1,2$. Для этого упорядочим значения c–го признака ($c \in I$) по возрастанию

$$r_{c_1}, r_{c_2}, \dots, r_{c_m}. \qquad (1)$$

Согласно определяемого ниже критерия последовательность (1) разбивается на $\tau_c, (\tau_c \geq 2)$ непересекающихся интервалов $\left[r_{c_u}, r_{c_v}\right]^i$, $1 \leq u, u \leq v \leq m$, $i = \overline{1, \tau_c}$. Значения, лежащие в интервале $\left[r_{c_u}, r_{c_v}\right]^i$, далее могут рассматриваться как градация номинального признака.

Пусть $d_t^i(u,v)$, $d_{3-t}^i(u,v)$, – количество представителей соответственно классов K_t, K_{3-t} в интервале $\left[r_{c_u}, r_{c_v}\right]^i$. Для рекурсивной процедуры выбора значений r_{c_u}, r_{c_v} используется критерий [4]

$$\left|\frac{d_t^i(u,v)}{|E_0 \cup K_t|} - \frac{d_{3-t}^i(u,v)}{|E_0 \cup K_{3-t}|}\right| \to \max. \qquad (2)$$

Границы первого интервала $\left[r_{c_u}, r_{c_v}\right]^1$ на последовательности (1) вычисляются по максимуму критерия (2). Аналогичным образом определяются границы для $\left[r_{c_u}, r_{c_v}\right]^p, p > 1$ на значениях (1), не вошедших в $\left[r_{c_u}, r_{c_v}\right]^1, \dots, \left[r_{c_u}, r_{c_v}\right]^{p-1}$. Критерием останова процедуры служит покрытие всех значений (1) непересекающимися интервалами.

Обозначим через

$$\eta_{1i}(t) = \frac{d_t^i(u,v)}{|E_0 \cup K_t|}, \eta_{2i}(t) = \frac{d_{3-t}^i(u,v)}{|E_0 \cup K_{3-t}|}$$

результаты оптимального разбиения по (2) для каждого интервала $\left[r_{c_u}, r_{c_v}\right]^i$, $i = \overline{1, \tau_c}$. Количественно доминирование выражается через значения функции принадлежности $f_t(i) \in [0,1]$ класса $K_t, t = 1,2$.

Значение функции принадлежности c–го признака к K_1 по интервалу $\left[r_{c_u}, r_{c_v}\right]^i$ определим как

$$f_1(i) = \frac{\eta_{1i}}{\eta_{1i} + \eta_{2i}}. \qquad (3)$$

С учётом того, что $f_t(i)=1 - f_{3-t}(i)$, $t=1,2$, устойчивость признака по множеству интервалов разбиения вычисляется как

$$U(c) = \frac{1}{m} \sum_{\{[r_u, r_v]^i\}} \begin{cases} f_t(i)(v-u+1), f_t(i) > 0.5, \\ (1 - f_t(i))(v-u+1), f_t(i) < 0.5, \end{cases} \qquad (4)$$

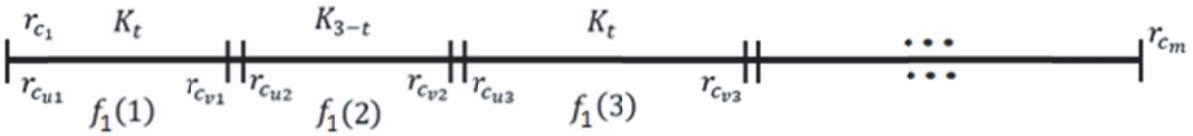

Рис. 1. Разбиение упорядоченных значений признака на интервалы

и выражает степень однородности (не перемешанности) значений с-го признака объектов в границах интервалов доминирования, определяемых по (2,3). Если (в идеале) в границах интервалов лежат значения признака одного класса, то $U(c) = 1$.

Визуальная интерпретация границ интервалов, полученных по (2), показана на рис. 1, где ($u1$,$v1$), ($u2$,$v2$),… – индексы упорядоченной последовательности (1). Нетрудно заметить, что не существует двух соседних интервалов, в которых доминировали представители одного класса.

Рассмотрим модификацию критерия (4) для случая наличия пропусков в данных. С учётом пропусков в данных критерий (4) примет вид

$$\left| \frac{d_t^i(u,v)}{T_p^c} - \frac{d_{3-t}^i(u,v)}{T_{3-p}^c} \right| \to max, \quad (5)$$

где T_p^c, T_{3-p}^c количество значений признака $x_c \in X(n)$ без пропусков у объектов E_0 соответственно из классов K_p и K_{3-p}. Естественным условием для реализации (5) является:

– число различных значений признака больше или равно 2;

– значения $T_p^c > 0, T_{3-p}^c > 0$.

С учётом пропусков в данных значение устойчивости (см. (4)) будет выглядеть так

$$U(c) = \frac{1}{\mu} \sum_{\{[r_u, r_v]\}} \begin{cases} f_t(i)(v-u+1), f_t(i) > 0.5, \\ (1-f_t(i))(v-u+1), f_t(i) < 0.5, \end{cases} \quad (6)$$

где $\mu = T_p^c + T_{3-p}^c$.

Примером формирования латентного признака из двух исходных, один из которых измеряется в количественной, а другой в номинальной шкале, может быть следующий. Пусть $x_i, x_j \in X(n), i \in I, j \in I$ и признак x_i имеет 2 градации. Тогда для получения латентного признака в виде произведения $x_i x_j$ значения признака x_i нужно выбирать из {-1,1}.

Разбиения на интервалы по (2) и (5) дают возможность для наглядного представления знаний в виде дизъюнкций элементарных конъюнкций. Элементарные конъюнкции нужны для проверки принадлежности значения признака к одному из интервалов. Запись правила для отнесения объекта классу K_t, t=1,2 может иметь вид: $a_1 \leqslant x_i \leqslant b_1$ or $a_2 \leqslant x_i \leqslant b_2$ or ….or $a_{\eta-1} \leqslant x_i \leqslant b_{\eta-1}$, где $a_j, b_j, j \in \{1, \eta\}$ – границы интервалов, η – число непересекающихся интервалов.

Значения устойчивости по (4) или (6) служат индикатором для использования разбиения на интервалы в качестве нового знания. Рекомендуется считать результаты анализа новым знанием при значении устойчивости из [0.9;1] и числе интервалов не больше 4.

2. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Для вычислительного эксперимента с целью поиска скрытых закономерностей были использованы данные Statlog [5] из UCI Machine Learning Repository. В Statlog содержатся данные сегментации изображений, которые разделены на семь классов (кирпич, небо, листва, цемент, окно, дорога, трава). Экземпляры (объекты) были случайно отобраны из базы данных открытых изображений. Каждый экземпляр представляет собой область из пикселей размера 3x3, количество экземпляров 2310. Для описания объектов выборки использовались 19 количественных признаков. Часть признаков получена по значениям интенсивности цветов от RGB генератора.

При проведении эксперимента выбирался один класс объектов изображений «кирпич» (K_1), все остальные объекты считались принадлежащими классу K_2. Результаты разбиения на интервалы по (2) и устойчивости по (4) приведены в табл. 1.

По результатам из табл. 1 устойчивость по (4) больше 0.9 у признаков 13, 18, 19. Согласно рекомендациям из п.1, именно эти признаки и границы их интервалов целесообразно использовать в качестве нового знания об объектах класса K_1 «кирпич», например, при формировании if …then правил в базах знаний. Полученное значение $U(3)$=0 объясняется тем, что не существует интервалов (для признака region-pixel-count) в которых по (2) доминируют представители одного из двух классов.

Очевидно, что рекомендации из п. 1 не могут быть ограничены исходными признаками в описании объектов классов. Дополнительные возможности для поиска скрытых закономерностей появляются при использовании в качестве исходных данных значений латентных признаков, синтезированных из исходных по правилам иерархической агломеративной группировки [3].

ЗАКЛЮЧЕНИЕ

Разбиение признаков на непересекающиеся интервалы и оценка его устойчивости разбиения служат хорошим средством для поиска

Таблица 1. Результаты разбиения на интервалы при выборе в качестве класса K_1 изображения «кирпич»

№	Название признака	Границы Интервалов	Значение функции (2) принадлежности к K_1	Устойчивость разбиения по (4)
1	region-centroid-col (столбец центрального пикселя области)	[1, 151]	0.5987	0.6557
		[152, 254]	0.2539	
2	region-centroid-row (строка центрального пикселя области)	[11, 50]	0.1533	0.7889
		[51, 149]	0.6607	
		[150, 251]	0	
3	region-pixel-count (количество пикселей в области = 9)	Нет	0	0
4	short-line-density-5 (результаты алгоритма экстракции линии, контраст, меньше или равный 5)	[0, 0]	0.4863	0.5222
		[0.1111, 0.3333]	0.5856	
5	short-line-density-2 (результаты алгоритма экстракции линии, контраст больше 5)	[0,0]	0.5089	0.5214
		[0.1111, 0.2222]	0.1714	
6	vedge-mean (измерение контраста по горизонтали используется как детектор вертикального края)	[0,0.2777]	0.1923	0.6116
		[0.2778, 0.6111]	0.750769	
		[0.6111, 29.2222]	0.4305	
7	vegde-sd (см. 6)	[0, 0.0333]	0.2857	0.6181
		[0.0333, 0.4333]	0.6797	
		[0.4333, 991.718]	0.4102	
8	hedge-mean (измеряется контраст вертикально смежных пикселей, используется для определения горизонтальной линии)	[0, 0.3333]	0.1046	0.6259
		[0.3333, 2.9444]	0.5662	
		[3, 44.7222]	0.2434	
9	hdge-sd (см. 8)	[-1.5e-008, 0.0296]	0	0.5981
		[0.0296, 0.4444]	0.6661	
		[0.4554, 1386.33]	0.4406	
10	intensity-mean (среднее значение интенсивности: среднее по области (R + G + B) / 3)	[0, 3.8889]	0.0179	0.8860
		[3.9259, 28.6296]	0.7443	
		[28.7407, 143.444]	0	
11	rawred-mean (среднее значение по области значения R)	[0,5.3333]	0.0956	0.8903
		[5.4444, 26.1111]	0.7685	
		[26.3333, 137.111]	0	
12	rawblue-mean (среднее значение по области значения B)	[0, 4.6667]	0.0453	0.8525
		[4.7778, 36.2222]	0.7207	
		[36.3333, 150.889]	0.0298	
13	rawgreen-mean (среднее значение по области значения G)	[0, 1.6667]	0	0.9103
		[1.7778, 20.6667]	0.7794	
		[20.7778, 142.556]	0.0104	

Таблица 1. Результаты разбиения на интервалы при выборе в качестве класса K_1 изображения «кирпич» (окончание)

14	exred-mean (избыток красного: (2R - (G + B)))	[-49.6667, -5.6667]	0.0790	0.8952
		[-5.5556, 7.2222]	0.8327	
		[9.8889, 9.8889]	0	
15	exblue-mean (избыток синего: (2B - (G + R)))	[-12.4444, 0.5556]	0.0316	0.8365
		[0.6667,23]	0.7494	
		[23.1111, 82]	0.1342	
16	exgreen-mean (избыток зеленого: (2G - (R + B)))	[-33.8889, -19.8889]	0.0933	0.8148
		[-19.7778, -6.3333]	0.6918	
		[-6.2222, 24.6667]	0.0441	
17	value-mean (среднее значение: трехмерное нелинейное преобразования RGB)	[0,5.3333]	0	0.8588
		[5.4444, 36.2222]	0.7230	
		[36.3333, 150.889]	0.0298	
18	saturatoin-mean (среднее значение насыщенности нелинейного преобразования RGB)	[0, 0.3679]	0.0052	0.9034
		[0.3688, 0.6170]	0.8057	
		[0.6176, 1]	0.1699	
19	hue-mean (среднее значение оттенка нелинейного преобразования RGB)	[-3.0442, -1.8905]	0.0190	0.9825
		[-1.8884, -0.5709]	0.9716	
		[-0.0049, 2.9125]	0	

скрытых закономерностей в данных. Обнаруженные закономерности являются источником нового знания в предметных областях.

СПИСОК ЛИТЕРАТУРЫ

1. Обучение распознаванию образов без переобучения / *Н.Г. Загоруйко, О.А., Кутненко А.О. Зырянов, Д.А. Леванов* // Машинное обучение и анализ данных. 2014. Т. 1. № 7. С. 891-901.
2. *Дуда Р., Харт П.* Распознавание образов и анализ сцен. Мир. 1976. – 512 с.
3. *Саидов Д.Ю.* Информационные модели на основе нелинейных преобразований признакового пространства в задачах распознавания: дис. ... докт. философии по физ.-мат. наукам. Ташкент, 2017. 104 с.
4. *Згуральская Е.Н.* Посик закономерностей по значениям колличественных признаков с помощью детерминистических критериев разбиения на интервалы // Междисциплинарные исследования в области математического моделирования и информатики. Материалы 3-й научно-практической internet-конференции. г. Тольяти 2014. С. 199-203.
1. Data & Knowledge Engineering 44 (2003) 109–138. UCI repository of machine learning databases. URL: http://archive.ics.uci.edu/ml/datasets/Statlog (дата

SUSTAINABILITY OF DIVIDING DATA IN INTERVALS IN THE PROBLEMS OF RECOGNITION AND SEARCHING FOR HIDDEN LAWS

© 2018 E.N. Zguralskaya

Institute of Aviation Technology and Management
of Ulyanovsk State Technical University

A great role for the improvement of digital technologies in the scientific, industrial and social spheres has the search for new knowledge contained in databases and data warehouses in the form of hidden patterns. In this paper, in order to identify hidden patterns in the recognition of objects, the method of splitting the characteristic values into disjoint intervals is considered. As a quality criterion, it is proposed to use the value of the stability indicator for dividing the original and latent features into intervals.

Keywords: hidden patterns, the stability of the division into intervals, data mining.

Ekaterina Zguralskaya, Senior Lecturer at the Aircraft Department. E-mail: iatu@inbox.ru