УДК 519.95

АЛГОРИТМ ВЫБОРА ОПТИМАЛЬНЫХ ГРАНИЦ ИНТЕРВАЛОВ РАЗБИЕНИЯ ЗНАЧЕНИЙ ПРИЗНАКОВ ПРИ КЛАССИФИКАЦИИ

© 2012 Е.Н. Згуральская

Институт авиационных технологий и управления
Ульяновского государственного технического университета

Поступила в редакцию 02.11.2012

Предлагается численный алгоритм выбора оптимальных границ интервалов разбиения значений признаков классифицированных объектов. Алгоритм инвариантен к масштабам измерений, может быть использован при поиске латентных (явно не измеримых) признаков в базах данных для моделирования процесса интуитивного принятия решений.
Ключевые слова: разбиение на интервалы, оптимальные значения границ интервалов, оценка сложности алгоритма.

ВВЕДЕНИЕ

Разбиение значений количественных показателей на интервалы широко применяется в различных алгоритмах анализа данных. В прикладной статистике значения количественных признаков, как правило, разбивается на заранее заданное число равных интервалов. Примером тому служит построение гистограмм, децильного и процентильного распределений.

Задача разбиения на интервалы рассматривалась и в теории распознавания образов с учителем. В [1] описан метод, реализация которого основывается на предположениях о законе распределения и числе интервалов. Метод является эвристическим, для разбиения на интервалы используется мера неопределённости принадлежности объекта к тому или иному классу энтропии, допускается отсутствие разбиения.

Использование численных методов оптимизации позволяет подбирать параметры модели, при которых алгоритмы распознавания допускают наименьшее число ошибок на заданной обучающей выборке. Метод, осуществляющий подгонку моделей распознавания и прогнозирования под выборку, получил название минимизации риска [2]. Увеличение сложности модели не всегда является благом, так как “оптимальные” алгоритмы начинают хорошо подстраиваться под конкретные данные, в том числе под измерения обучающей выборки и погрешность самой модели.

В теории искусственных нейронных сетей (ИНС) сложность модели распознавания выражается через способность к обобщению. Требуется, чтобы алгоритмы ИНС не только хорошо решали задачу на обучении, но и были способны также хорошо принимать решение на объектах, которые они не видели в процессе обучения. Этим целям служат разработки новых методов интеллектуального анализа данных, позволяющих получать новые знания о решаемой задаче и использовать их, в том числе, и для повышения точности алгоритмов ИНС [3] для произвольных допустимых объектов.

Задача разбиение на интервалы значений признаков классифицированных объектов в [3] сформулирована как детерминистическая. В основе критерия метода лежит проверка гипотезы “*Существует такое разбиение, при котором каждый интервал содержит все значения признака объектов одного класса*”. Очевидно, что при проверке число интервалов должно быть равно числу классов объектов.

Истинность, указанной выше гипотезы, означает, что между интервалами значений количественных признаков и классами объектов существуют взаимно-однозначное соответствие. На практике интерес представляет ответ на вопрос: Насколько истинны утверждения гипотезы на реальных данных? Универсальной и легко интерпретируемой мерой истинности служат значения в интервале $[0,1]$. Концы интервала $[0,1]$ определяют оппозицию: значения признака неразличимы – значения признака различимы до уровня взаимно-однозначного соответствия интервалов и классов объектов.

Описываемый в работе алгоритм инвариантен к масштабам измерений, может быть использован при:

- поиске латентных (явно не измеримых) признаков в базах данных для моделирования процесса интуитивного принятия решений;

- преобразовании значений количественных признаков в номинальные с минимальной потерей информации;

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

- отборе информативных наборов разнотипных признаков.

Для уменьшения объёма вычислений предлагается проводить предобработку данных. Даётся оценка комбинаторной сложности алгоритма без использования предобработки и при её использовании.

1. ПОСТАНОВКА ЗАДАЧИ И МЕТОД РЕШЕНИЯ

Рассматривается задача распознавания в стандартной постановке. Считается, что задано множество объектов $E_0 = \{S_1,...,S_m\}$, содержащее представителей l непересекающихся классов $K_1,...,K_l$. Описание объектов производится с помощью набора из n разнотипных признаков $X_n = (x_1,...,x_n)$, δ из которых измеряются в номинальной шкале, $n-\delta$ в интервальной шкале. Считается, что задан критерий $F(*)$ для разбиения значений количественного признака на непересекающиеся интервалы. Требуется определить значения границ l интервалов при $F(*) \to extr.$

Обозначим через I, J множество номеров соответственно количественных и номинальных признаков в описании допустимых объектов, $|I| + |J| = n$. Упорядоченное множество значений признака $x_j, j \in I$ разобьём на непересекающиеся интервалы $(c_{2k-1}, c_{2k}], c_{2k-1} < c_{2k}$, $k = \overline{1,l}$, каждый из которых считается градацией номинального признака.

Пусть u_i^p - множество значений признака $x_j, j \in I$ класса K_i в интервале $(c_{2p-1}, c_{2p}]$, $A = (a_0,...,a_l), a_0 = 0, a_l = m$, a_p – порядковый номер элемента упорядоченной по возрастанию последовательности $r_{j1},...,r_{jm}$ значений x_j из E_0, определяющий правую границу интервала $c_{2p} = r_{a_p}$.

Критерий

$$\left(\frac{\sum_{p=1}^{l}\sum_{i=1}^{l} u_i^p(u_i^p-1)}{\sum_{i=1}^{l}|K_i|(|K_i|-1)}\right)\left(\frac{\sum_{p=1}^{l}\sum_{i=1}^{l} u_i^p(m-|K_i|-\sum_{j=1}^{l} u_j^p+u_i^p)}{\sum_{i=1}^{l}|K_i|(m-|K_i|)}\right) \to \max_{\{A\}} \quad (1)$$

позволяет вычислять оптимальные значения границ интервалов $\{(c_{2p-1}, c_{2p}]\}$ и использовать их для определения градаций количественного признака в номинальной шкале измерений. Процесс преобразования при этом оказывается неразрывным от классификации, вводимой на множестве объектов обучения, и может быть реализован с учётом пропусков в данных.

Основные затраты вычислительных ресурсов при нахождении экстремума (1) приходится на вычисление $\{u_i^p\}$. Максимальное число непересекающихся интервалов при разбиении упорядоченной последовательности $r_{j1},...,r_{jm}$ при числе классов l равно

$$\psi(l,m) = \begin{cases} 2(m-1), l = 2, \\ 2\prod_{k=3}^{2(l-1)}(m-2l+k), l > 2. \end{cases} \quad (2)$$

Количество операций (сложность алгоритма) для подсчёта $\{u_i^p\}$ определяется по среднему числу проверок условий вхождения значения признака в один из l интервалов (по две в интервале) и операции суммирования с 1

$$F(l,m) = \left(2m\left(\frac{1+l}{2}\right)+m\right)\psi(l,m) = m(2+l)\psi(l,m).$$

Для уменьшения комбинаторной сложности вычислений предлагается воспользоваться предобработкой данных. Суть предобработки заключается в формированиии по упорядоченной последовательности $r_{j1},...,r_{jm}$ целочисленной матрицы вида

$$D = \begin{pmatrix} d_{10}d_{11}\ldots d_{1m} \\ \ldots\ldots\ldots\ldots \\ d_{l0}d_{l1}\ldots d_{lm} \end{pmatrix}, \quad (3)$$

в которой индекс столбца элемента $d_{pi}, p = \overline{1,l}, i = \overline{1,m}$ соответствует объекту $S \in E_0$ со значением признака r_{ji}. Элементы матрицы (3) вычисляется как

$$d_{pi} = \begin{cases} 0, i = 0, \\ d_{p,i-1} + g(p,i), i > 0, \end{cases} \text{где } g(p,i) = \begin{cases} 1, S \in K_p, \\ 0, S \notin K_p. \end{cases}$$

Число представителей u_t^p класса $K_p, p = \overline{1,l}, t = \overline{1,l}$ в интервале $[c_1, c_2]$ при $t = 1$ и $(c_{2t-1}, c_{2t}]$ при $t > 1$, левые и правые границы которых соответствуют индексам $\eta = a_{t-1}$, $\nu = a_t$, $c_{2t-1} = r_{j\eta}$, $c_{2t} = r_{j\nu}$, определяется как

$$u_t^p = d_{p\nu} - d_{p\eta}. \quad (4)$$

Сложность алгоритма вычисления $\{u_i^p\}$ по (3),(4) не превышает $l\psi(l,m)$. Эта оценка сложности может быть понижена при наличи пропусков в данных и повторяющихся значений.

Благодаря использованию (4) по матрице (3) стало возможным вычисление (1) для интервалов и весов латентных (явно не измеримых) признаков. Под весом здесь понимается оптимальное значение критерия (1). Примерами латентных признаков может служить $x_i x_j, x_i x_j^{-1}$, где $x_i, x_j \in X$ и $i, j \in I$. На практике латентные признаки часто используются в форме различных

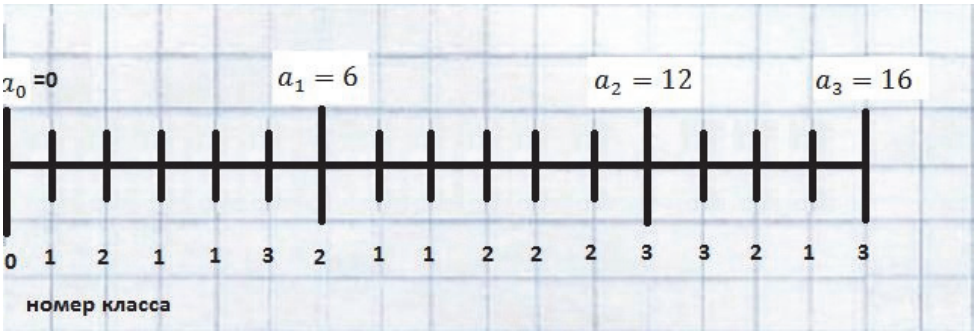

Рис. 1. Визуальная демонстрация алгоритма

индексов. Например, в медицине это индекс массы тела, индекс Кердо. Высокие значения весов латентных признаков (как правило, ближе или равные 1) служать основанием для построения моделей интуитивного принятия решений.

Вес признака по критерию (1) содержит в себе важную информацию об его информативности. Однако при отборе информативных наборов признаков нельзя полностью полагаться только на их упорядочение по значениям весов, то есть руководствоваться принципом "чем больше вес, тем признак более информативный в наборе". В расчёт идёт и такой фактор как взаимная коррелированность признаков. Такая задача рассматривалась в [4], где исследовался вопрос отбора наборов информативных разнотипных признаков и их влияние на эффективность реализации искусственных нейронных сетей.

2. ТЕСТОВЫЙ ПРИМЕР

Визуальная демонстрация алгоритма разбиения на интервалы несовпадающих значений количественного признака по критерию (1) при m = 16, числе классов l =3 и мощности классов $|K_1| = 6,\quad |K_2| = 6,\quad |K_3| = 4$ с использованием результатов предобработки (5) показана на рис. 1.

$$D=\begin{pmatrix} 0 & 1 & 1 & 2 & 3 & 3 & 3 & 4 & 5 & 5 & 5 & 5 & 5 & 5 & 6 & 6 \\ 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 3 & 4 & 5 & 5 & 5 & 6 & 6 & 6 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 3 & 3 & 3 & 4 \end{pmatrix}. \quad (5)$$

Согласно (2) максимальное число вариантов разбиения значений признака на интервалы $\psi(3,16) = 2\times13\times14 = 364$, сложность алгоритма без предобработки $F(3,16) = 16\times(2+3)\times364 = 29520$ с предобработкой (учитывая вычисление (3)) $3\times364 + 3\times16 = 1140$. Для варианта разбиения, указанного на рис. 1, получим u_1^1=3, u_1^2=2, u_1^3=1, u_2^1=2, u_2^2=3, u_2^3=1, u_3^1=1, u_3^2=1, u_3^3=2 и значение критерия (1) равное 0,2146.

Другие возможные варианты разбиения на интервалы представлены в табл. 1.

Таблица 1. **Варианты разбиения на интервалы**

№ п/п	a_1	a_2	Значение критерия (1)
1	1	2	0,1944
2	2	8	0,3452(оптим.)
3	6	12	0,2146

Очевидно, что при оптимальном разбиении ($a_1 = 2$ и $a_2 = 8$) нет ни одного интервала, содержащего все значения признака объектов одного класса.

Номера интервалов оптимального разбиения по (1) можно рассматривать как градации при преобразовании значений количественного признака в номинальные. Такое преобразование использовалось для поиска информативных наборов разнотипных признаков с максимально выраженной независимостью в [4].

3. ВЫВОДЫ

Численный алгоритм выбора оптимальных границ интервалов может быть использован при интеллектуальном анализе данных для преобразования количественных признаков в номинальные с минимальной потерей информации, при упорядочении разнотипных признаков по отношению сложности алгоритмов, выражаемой через способность корректно распознавать объекты фиксированной выборки с минимальной затратой вычислительных ресурсов.

СПИСОК ЛИТЕРАТУРЫ

1. *Вапник В.Н.* Алгоритмы и программы восстановления зависимостей. М.: Наука, 1984. 816 с.
2. *Вапник В.Н.* Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979.447с.
3. *Игнатьев Н.А., Мадрахимов Ш.Ф.* О некоторых способах повышения прозрачности нейронных сетей // Вычислительные технологии. 2003. Т. 8. № 6. С. 31-37.
4. *Згуральская Е.Н.* Выбор информативных признаков для решения задач классификации с помощью искусственных нейронных сетей // Нейрокомпьютеры: разработка, применение. 2012. №2. С. 20-26.

THE ALGORITHM OF DETERMINING OF THE OPTIMAL PARTITION BOUNDARIES ATTRIBUTE VALUES INTERVALS FOR THE CLASSIFICATION

Institute of Aviation Technology and Management
of Ulyanovsk State Technical University

A numerical algorithm for choosing the optimal boundaries of partition values intervals of the classified objects attributes is represented. The algorithm is invariant to the scale of the measurement and can be used for finding latent (not measurable clearly) criteria in the database for modeling of the intuitive decision-making process.
Keywords: partition into intervals, the optimal values of the boundaries of intervals, the evaluation of the algorithm.

Ekaterina Zguralskaya, Senior Lecturer at the Aircraft Structure Department. E-mail: e_ignateva@rambler.ru