

粗糙模型

VPRS

增量式归纳学习

人工智能

(14)

55-58,61

计算机科学 1999 Vol. 26 No. 1

基于可变精度粗集模型的增量式归纳学习

Incremental learning based on Variable Precision Rough Set Model

李莉

TP18

(西南师范大学计算机科学系 重庆 400715)

Abstract Symbolic inductive learning methods such as induction of decision trees and rule induction methods are used in knowledge discovery from databases (KDD). However, most approaches cannot classify problems involving uncertain or imprecise information. In this paper, we discuss variable precision rough set model based approach to induce rules that aim at classification problems incrementally.

Keywords Rough set, Incremental learning, KDD, Variable precision rough set model

知识获取^[1]是知识信息处理系统的关键问题之一,特别是计算机与网络技术的飞跃发展使得各个领域的信息量急剧增加,90年代以来,人们把目光转向了一个新的研究领域——知识发现(Knowledge Discovery),或数据库知识发现(Knowledge Discovery from Databases, KDD)。今天,基于Web的分布式知识发现更为KDD展示了广阔的应用前景。

在KDD'96国际学术会议上Fayyad U^[1]等人对KDD作出了定义:KDD是识别出存在于数据库中有效的、新颖的、具有潜在效用的乃至最终可理解的非凡性过程。其中,有效性强调过程的计算机实现;非凡性是指无法利用一般数据库技术实现。这次会议上,确定知识发现的主要步骤如下:

- 1)理解领域知识和相关的先验知识,确定系统目标;
 - 2)创建相关的目标数据集;
 - 3)数据整理和预处理;
 - 4)数据约简和投影;
 - 5)选择与1)所确定目标相应的数据挖掘任务;
 - 6)选择发现算法,即用于搜索数据中模式的方法;
 - 7)数据挖掘,实施如数据聚类、分类或预测模型知识发现、数据总结、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等发现任务;
 - 8)用可视化解释得到的模式,并重复1)至7);
 - 9)巩固得到的知识。
- 近年来基于粗集理论^[2]的方法已成为知识发现

的主要方法之一。粗集模型的中心任务就是分类,但是,基本的粗集模型不能形成一定错误分类率或一定程度的不确定性的分类,而实际应用中,局部的不能正确分类的规则却可以提供一种对大部分可利用数据的准确分类,而且,KDD经常需要将过去发现的知识用于新的范例,以获取一般性的结论。基于此,Ziarko W.提出了可变精度粗集模型^[1](Variable Precision Rough Set Model,简称VPRS),以不大于给定错误分类率为目的来产生具有一定不确定性或误分类率的近似分类。本文在讨论了可变精度粗集模型 γ 程度上的 β -相关的基础上,引入了基于规则的归纳学习方法。

1. 可变精度粗集模型

1.1 粗集理论

粗集理论由Pawlak Z.提出,其主要思想就是在保持分类能力不变的前提下,通过知识约简,导出概念的分类规则。粗集理论认为知识具有粒度(granularity),而且不精确性主要由于粒度太大而引起,解决的方法之一就是形成知识库的偏序结构,并通过运算,求出最小约简,形式化条件属性和决策属性之间的最小依赖关系,而不必针对所有的条件属性集。粗集理论为处理含噪声、不精确或不完整的数据分类问题,提供了严密的数学工具。

近似空间 $A = \langle U, R \rangle$, U 是对象的非空有限集,称为论域, $R \subseteq U \times U$ 是 U 上的等价关系,称关系 R 的等价类 $[x]_R$ 为 A 中的基本集,记为 $R^* = \{E_1, E_2, \dots, E_n\}$,也把等价关系 R 称为不可辨认性(indis-

cernibility)关系。粗集理论主要通过 R 上的基本集, 利用部分论域集合而达到分类的目的。设 $X \subseteq U$, 在 U 中定义 X 的上近似和下近似为:

$$L_R(X) = \{x \in U \mid [x]_R \subseteq X\}$$

$$U_R(X) = \{x \in U \mid [x]_R \cap X \neq \emptyset\}$$

其中 $[x]_R = \{y \in U \mid xRy\}$, R 为等价关系。

$L_R(X)$ 可理解为 A 中基本集的联合, 也称为正区域。图 1 形象地描述了上近似和下近似两个概念。其中,

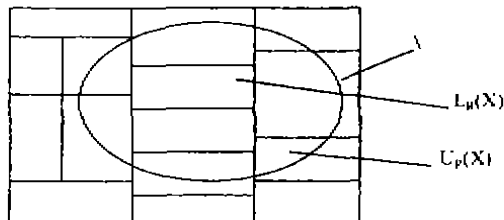


图 1 上近似和下近似图示

大的矩形框表示论域, 小的矩形框表示等价类亦即划分。

知识表示系统也称信息系统^[4], 由四元组 $S = (U, A, V, f)$ 组成, 其中 U 为对象的集合, A 为对象属性的集合, 且 $A = C \cup D$, 即属性集合包含条件属性集和决策属性集。 V 是属性域的并, 即 $V = \bigcup V_a$, 对 $a \in A$, V_a 定义为属性 A 的域, 而称 f 为信息函数, 定义为: $f(x, d) \rightarrow V_a$, 其中 $x \in U, d \in D, d \in A, V_a \in V$ 。定义 $IND(C)$ 和 $IND(D)$ 是 U 上关于属性 C 和 D 的两个不可辨认关系, 并满足: $x, y \in U$, 且 x, y 关于 C 等价, 即对所有的 $a \in C, (x, y) \in IND(C)$ 当且仅当 $f(x, a) = f(y, a)$ 。信息系统可方便地表示为知识表的形式。知识表有时也称为数据表、属性值表或条件动作表。它由行、列以及相对应的值表示。本文中, 我们主要讨论 C 和 D 两大类属性即条件属性和决策属性之间的关系。

例 1 $X_1 = \{x_1, x_2, x_3\}, X_2 = \{x_3, x_2, x_4, x_7\}, X_3 = \{x_1, x_7, x_3\}, Y = \{x_1, x_2, x_4, x_8\}$

由于粗集理论严格按等价类来分类属性, 因而它的分类是精确的, 亦即“包含”或“不包含”, 而没有某种程度上的“包含”或“属于”。但实际应用中却有这样的例子存在, 比如说, “计算机系四年级的学生, 95% 成绩优秀”, 则从数据库中抽取的这条规则将“计算机系四年级的学生”与等级“优秀”联系起来, 而且具有较高的可信度。反过来, 如果预测“优秀”, 且告知是“计算机系四年级的学生”, 则不正确的可

能性就非常小。同样对于例 1, $X_1 \subseteq Y, X_2 \subseteq Y, X_3 \subseteq Y$ 等包含关系就不成立, 如果放宽条件, 引入一小正数 $\beta (0 \leq \beta < 0.5)$, 将会得到 $\{X_1, X_2, X_3\}$ 与 Y 之间的一定程度上的关系。

1.2 可变精度粗集模型

可变精度粗集模型^[3]是粗集模型的扩充, 它是在基本粗集模型的基础上引入了 $\beta (0 \leq \beta < 0.5)$, 即允许一定程度的错误分类率 (misclassification) 存在, 这一方面完善了近似空间的概念, 另一方面也有利于用粗集理论从认为不相关的数据中发现相关数据。当然, VPRS 的主要任务是解决属性间无函数或不确定 (non-deterministic) 关系的数据分类问题。当 $\beta = 0$ 时, 粗集模型是 VPRS 的一个特例。

VPRS 讨论了属于“多少”程度的“包含”关系。设 X, Y 表示集合:

$$C(X, Y) = \begin{cases} 1 - \text{card}(X \cap Y) / \text{card}(X) & \text{card}(X) > 0 \\ 0 & \text{card}(X) = 0 \end{cases} \quad (1)$$

定义 1 $C(X, Y)$ 为错误分类率, 有

$$Y \supseteq X \Leftrightarrow C(X, Y) \leq \beta \quad (2)$$

其中 $\text{card}(X)$ 是集合的基数。引入 $\beta (0 \leq \beta < 0.5)$ 后, 由定义 1, 例 1 有下式成立:

$$Y \supseteq X \Leftrightarrow C(X, Y) = 0, Y \supseteq^{0.35} X_1, Y \supseteq^{0.25} X_2, Y \supseteq^{0.33} X_3$$

设 U 为论域, $X \subseteq U, R$ 为 U 上的等价关系, E 为等价类, $R = \{E_1, E_2, \dots, E_s\}, E_i = [x]_R$, 定义 β -上近似:

$$\begin{aligned} \bar{R}_\beta X &= \bigcup \{E \in R^* \mid C(E, X) < 1 - \beta\} = \\ &\bigcup \{E \in R^* \mid X \supseteq^{1-\beta} E\} \end{aligned} \quad (3)$$

定义 β -下近似:

$$\begin{aligned} R_\beta X &= \bigcup \{E \in R^* \mid C(E, X) \leq \beta\} = \\ &\bigcup \{E \in R^* \mid X \supseteq^\beta E\} \end{aligned} \quad (4)$$

(4) 式也称 β -正区域, 记为 $\text{POSR}_\beta(X)$ 。

定义 β -边界域:

$$\text{BNR}_\beta(X) = \bigcup \{E \in R^* \mid \beta < C(E, X) < 1 - \beta\} \quad (5)$$

定义 β -负区域:

$$\text{NEGR}_\beta(X) = \bigcup \{E \in R^* \mid C(E, X) \geq 1 - \beta\} \quad (6)$$

β -正区域可理解为将 U 中的对象属性以小于 β 的分类误差分类于 X 的集合, β -负区域相应理解为将 U 中的对象属性以小于 β 的分类误差分类于 X 的补集即 $(-X)$ 的集合。

对于 $\forall X, X \subseteq U$, 有 (7) 式成立:

$$\text{POSR}_\beta(-X) = \text{NEGR}_\beta(X) \quad (7)$$

如果 $\text{BNR}_\beta(X) = \emptyset$, 则

$$POS_{\beta}(X) \cup NEGR_{\beta}(X) = U \quad (8)$$

为描述集合 X 被近似空间 $A = (U, R)$ 中的基本元素近似分类的程度, 利用文献[2]准确程度因子, 定义 β -准确度 ($0 \leq \beta < 0.5$) 为:

$$\alpha(R, \beta, X) = \text{card}(\underline{R}_{\beta}X) / \text{card}(\overline{R}_{\beta}X) \quad (9)$$

当 β -准确度等于单位 1, 或 $\text{BNR}_{\beta}(X) = \Phi$, 即 (10) 式成立时, 称集合 X 为 β -不可辨认 (β -indiscernable):

$$\overline{R}_{\beta}X = \underline{R}_{\beta}X \quad (10)$$

定义 2 γ 程度上的 β -依赖为:

$$\gamma(P, Q, \beta) = \text{card}POS(P, Q, \beta) / \text{card}U \quad (11)$$

其中 $POS(P, Q, \beta) = \bigcup_{Y \in Q} IND(P, \beta)Y$ (12)

$IND(P)$ 定义为不可辨认的关系, 则称决策属性 Q 在程度 γ 上与条件属性 P 关于 β -依赖。从上述讨论可知, 引入 β ($0 \leq \beta < 0.5$) 参数后, 扩充了基本粗集理论, 更好地体现了数据分析中的数据相关性, 从而为获取近似决策规则奠定了基础。

定义 3 定义 $RED(P, Q, \beta)$ 为条件属性 P 相当于决策属性 Q 的 β -约简 (β -reduct) 或称近似约简。其中 $RED(P, Q, \beta) \subseteq P$, 而且有: 1) $\gamma(P, Q, \beta) = \gamma(RED(P, Q, \beta), Q, \beta)$; 2) 设存在另一个 $RED'(P, Q, \beta) \subseteq P$, 则有 $RED'(P, Q, \beta) \supseteq RED(P, Q, \beta)$ 成立。即 $RED(P, Q, \beta)$ 为最小约简集合。从 $RED(P, Q, \beta)$ 中删除任何属性, 都将使 1) 不成立。

2. 增量式学习

为便于描述整个过程, 以一个简单的数据表 1 为例。对于连续取值属性的离散化, 可利用等宽划分、等频划分以及 ChiMerge^[7] 等方法, 本文假定属性已经离散化处理。

表 1 简单数据表

U	颜色	大小	形状	类别
1	B	L	S	3
2	Y	S	C	2
3	B	L	S	3
4	Y	M	R	2
5	B	M	C	2
6	B	L	S	2
7	R	L	C	2
8	Y	M	R	3
9	B	M	C	3
10	Y	S	C	2
11	B	M	C	3
12	R	L	C	1
13	Y	M	R	1
14	B	M	C	1
15	Y	M	R	1

注: 颜色、大小、形状均分三种, 分别为: red, blue, yellow; small, large, medium; square, circle, rectangle.

两类属性 $P = \{\text{颜色, 大小, 形状}\}$, $Q = \{\text{类别}\}$, 划分 P^* 上有五个类别, $X_1 = \{1, 3, 6\}$, $X_2 = \{2, 10\}$, $X_3 = \{4, 8, 13, 15\}$, $X_4 = \{5, 9, 11, 14\}$, $X_5 = \{7, 12\}$, 划分 Q^* 上有三个类别, $Y_1 = \{12, 13, 14, 15\}$, $Y_2 = \{2, 4, 5, 6, 7, 10\}$, $Y_3 = \{1, 3, 8, 9, 11\}$ 。考虑新增一项:

16	R	L	C	1
----	---	---	---	---

如何在原有的基础上扩充知识, 正是我们下面要讨论的。

将相对于 P 和 Q 的划分分别记为 P^* , Q^* , $P^* = (X_1, X_2, \dots, X_n)$, $Q^* = (Y_1, Y_2, \dots, Y_m)$, m, n 为正整数, 并设定分类误差 β ($0 \leq \beta < 0.5$)。接着计算:

$$Y_i \overset{\beta}{\supseteq} X_j, i \in [1, n], j \in [1, m]$$

并定义:

$$\text{Set}P = \{X_j | X_j \overset{\beta}{\subseteq} Y_i, i \in [1, n], j \in [1, m]\}$$

$$\text{Set}Q = \{Y_i | Y_i \overset{\beta}{\supseteq} X_j, i \in [1, n], j \in [1, m]\}$$

$$\text{REDSet} = \{[RED(P, Q, \beta), N] | \gamma(RED(P, Q, \beta)) = \gamma(RED(P, Q, \beta), Q, \beta), N \text{ 为整数}\}$$

定义 TEMPSet 为属性的集合, 初始时空, 定义 oldattr_i 为 $RED(P, Q, \beta)$ 集合中各规则的范式, 随着新的观测值的加入, 整个过程如下:

1) 将添加的各属性项与 REDSet 集中的子集作比较, 若出现相对能接收的误差, 则结束返回。

2) 根据 P^* , Q^* 的定义, 划分等价类: ①若没有出现新的类别, 即 $X_{\text{NEW}} \in \text{Set}P$, $Y_{\text{NEW}} \in \text{Set}Q$, 则计算 $C(X_{\text{NEW}}, Y_{\text{NEW}})$, 并判定是否 $C(X_{\text{NEW}}, Y_{\text{NEW}}) \leq \beta$; ②其余情况, 则需计算 $C(X_{\text{NEW}}, Y_j)$, $j \in [1, m]$, 判定是否 $C(X_{\text{NEW}}, Y_{\text{NEW}}) \leq \beta$ 。

3) 根据公式 (11) 计算 $\gamma(P, Q, \beta)$ 的值,

4) 计算 $RED(P, Q, \beta)$, 使得 $\gamma(P, Q, \beta) = \gamma(RED(P, Q, \beta), Q, \beta)$ 。

5) 将 newattr_i [$RED(P, Q, \beta)$, 1] 的属性组合, (由析取和合取运算组成的范式, 同 oldattr_i) 与 REDSet 中的集合元素 oldattr_i 比较, (i 为属性集合的个数), 分两种情况: ① newattr_i $\cap$ oldattr_i = Φ , 则 TEMPSet = TEMPSet $\cup$ {newattr_i}, 并标记触发次数记等于 $N = 1$; ② newattr_i $\cap$ oldattr_i $\neq \Phi$, 除非两者相等, 标记触发次数 $N = N + 1$, 否则, REDSet =

REDSet $\cup\{\text{newattr}\}$,也相应标记触发次数 N 。

6)若 TEMPSet 集合中某一属性组合的触发次数 N 超过给定的阈值,才将其添加并入集合 REDSet 中,这样可有效防止环境因素的变化或带噪声数据的影响,只有当某一规则多次出现时,才考虑增加 REDSet 中的属性。如果经过一段时间的观察,发现 REDSet 中的某规则触发的概率很低,也可考虑将其删除,并入 TEMPSet 中。

7)重复上述步骤。

下面根据表 1 给出计算过程:设定 $\beta=0.34$,经计算有:

$$X_1 \stackrel{0.33}{\subseteq} Y_3 \text{ 和 } X_2 \stackrel{0.35}{\subseteq} Y_3$$

则 $\gamma(P,Q,0.33)=\text{card}(X_1 \cup X_2)/\text{card}(U)=5/15=1/3$;增加新项后, X_3 和 Y_1 发生变化, $X_3=\{7,12,16\}$, $Y_1=\{1,3,12,13,14,15,16\}$,重新计算 $C(X_3, Y_1)$, $C(X_3, Y_1)=1/3$,则有:

$\gamma(P,Q,0.33)=\text{card}(X_1 \cup X_2 \cup X_3)/\text{card}(U)=7/16 \geq 1/3$;由于例子的属性不多,约简后的规则集变化不大。

3. 最佳约简集

计算所有约简和计算一个最佳约简是 NP 难题^[5]。求最小约简集合的主要困难来源于没有一个选择关联属性的最佳标准。针对不同的问题,可以代价最小或路径最短等为目标,其它情况下,只好从数据表入手,并遵循:(1)选择属性最少的约简;(2)选

择属性取值组合最少的约简。

结束语 粗集理论作为研究不确定性、含糊性数据的重要工具,在 KDD 中日益显示出它的重要性。特别是基于 Web 的分布式信息的迅猛增加,人们更希望从浩如烟海的信息中提炼出自己感兴趣的数据,这一方面有赖于知识源的整理加工,另一方面也需要灵活方便的分析方法以及各种智能的工具,来代替烦琐的人工搜索过程。粗集理论及其扩展模型,都较好地解决了 KDD 中数据的不完整性、不确定性等问题,但是,由于粗集理论缺乏对数据源模糊性的处理;对于模糊概念的边界区域刻画太简单;不能提供一定容忍度(一定错误分类率)的分类方法等,因而仍需不断地与其它数学工具以及人工智能的方法相结合,从而开发客观现实世界蕴藏的丰富信息资源。

参考文献

- 1 Fayyad U, et al. Knowledge discovery and data mining: towards an unifying framework. In: Simoudis E, Han J-W, Fayyad U M Eds. Proc. of AAAI Workshop on KDD-96. Menlo Park, California: AAAI Press, 1996
- 2 Pawlak Z. Rough sets. Intl. J. of Computer and Information Science, 1982, 11(5)
- 3 Ziarko W. Variable precision rough set model. J. of computer and system sciences, 1993(46)

(下转第 61 页)

《计算机科学》参考文献表的编写格式

《计算机科学》采用顺序编码制。在文后参考文献表中,各条文献按在论文中的文献序号排列顺序,项目应完整,内容、著录格式与符号应正确。所用参考文献应列出主要者,尽量限制在 10 篇内。下面列出常见的几种编写格式:

1 专著

标引项顺序号 著者. 书名. 版本. 其他责任者. 出版地:出版者,出版年

2 论文集中析出的文献

标引项顺序号 作者. 题名. 见(In):编者. 文集名. 出版地:出版者,出版年. 在原文献中的位置

3 期刊中析出的文献

标引项顺序号 作者. 题名. 其他责任者. 刊名,年,卷(期):在原文献中的位置

4 学位论文

标引项顺序号 作者. 题名:[学位论文]. 保存地:保存者,年份

在参考文献中,用各种文字书写的姓名,一律姓在前名在后;外国人名可缩写为首字母(大写),但不加缩写点。

$$\neq, \tau(K')))))).$$

(2) 交($\cap$)

$$r \cap s = \{\tau \mid \exists \tau_1 \in r, \exists \tau_2 \in s (K' \in K(\tau_1(K') = \tau_2(K')) \wedge \tau = \{\min(\pi_{\tau_1}(u_i), \pi_{\tau_2}(u_i))/u_i : u_i \in U, i=1, 2, \dots, n\})\}$$

(3) 差($-$)

$$r - s = \{\tau \mid \tau \in r \wedge (\forall \tau' \in s (\exists K' \in K(\tau(K') \neq \tau'(K'))))\}$$

(4) 笛卡尔积($\times$)

设 s' 是模糊模式 $S = \{V_1 \wedge V_2 \times \dots \wedge V_m\}$ 的模糊关系, 则

$$r \times s' = \{(a_1, a_2, \dots, a_n, b_1, \dots, b_m) : a_i \in U, b_j \in V_j, 1 \leq i \leq n, 1 \leq j \leq m\}$$

(5) 选择(σ)

设 r' 是模糊模式 R 上的一个模糊关系, Q 是一个 k 元模糊谓词, 则选择操作定义如下:

$r = \sigma_Q(r')$, 其中 Q 与传统谓词的区别是“=”被置换成“ $\approx$ ”。

(6) 投影(π)

投影是使用并操作将传统投影合成一个关系。设 r 是模糊模式 R 上的一个模糊关系, 它的属性是 $A_1, A_2, \dots, A_m$, $S = \{A_{i_1}, A_{i_2}, \dots, A_{i_k}\}$ 是 R 的子集, 则 r 在 S 上的投影定义为:

$$\pi_{A_{i_1}, A_{i_2}, \dots, A_{i_k}}(r) = \bigcup \{S(A_{i_1}, A_{i_2}, \dots, A_{i_k})\}$$

(7) 自然连接($\bowtie$)

设 r 是模糊模式 $R = \{U_1 \times U_2 \times \dots \times U_m\}$ 上的一个模糊关系, s 是模糊模式 $S = \{V_1 \wedge V_2 \times \dots \wedge V_n\}$ 的一个模糊关系, $R \cap S = P$ 且 $P \neq \emptyset$, 则 r 和 s 的自然连接定义为:

$$r \bowtie s = \{\tau \mid \exists \tau_1 \in r, \exists \tau_2 \in s (\forall K' \in P(\tau_1(K') = \tau_2(K')) \wedge \tau(R-P) = \tau_1(R-P) \wedge \tau(P) = (\min(\pi_{\tau_1}(u_i), \pi_{\tau_2}(u_i))/u_i : u_i \in P) \wedge \tau(S-P) = \tau_2(S-P))\}. \text{ 其中 } i=1, 2, \dots, t, \text{ 这里 } t \text{ 是 } U_i \text{ 的模糊值域的维数。}$$

四、结论

上面, 我们已经给出了模糊的关系演算语言和关系代数操作的定义, 从这些定义来看, 显然可以得到如下定理:

定理 1 模糊的第一范式(1NF)在并、交、差、笛卡尔积、投影、选择和自然连接操作下是封闭的。

定理 2 如果 E 是一个模糊的 1NF 关系代数表达式, 则必定存在一个与之等价的安全模糊的 1NF 元组关系演算表达式 E' 。

定理 3 如果 E 是一个安全模糊的 1NF 元组关系演算表达式, 则存在一个与之等价的模糊的 1NF 关系代数表达式 E' 。

定理 4 对于每一个安全模糊的 1NF 元组关系表达式 E , 都存在一个与之等价的安全模糊域关系演算表达式 E' 。

定理 5 对于每一个安全的模糊域关系演算表达式 E , 都存在一个与之等价的模糊的 1NF 关系代数表达式 E' 。

说明: 这些定理的证明可以由传统的关系数据库理论中类似的定理的证明进行构造, 其差异是“=”被置换成“ $\approx$ ”。

参考文献

- 何新贵. 模糊数据库中的语义距离及模糊视图. 计算机学报, 1989, 10: 757~764
- Takahashi Y. Fuzzy Database Query Languages and Their Relational Completeness Theorem. IEEE Trans. on Knowledge and Data Engineering, 1993, 5 (1): 122~125
- Ullman J D. Principle of Database Systems. Computer Science Press, 1980
- Yu Hairong, Ramer A. A Combined Approach to Uncertain Data Analysis. Advance in Intelligent Data Analysis (IDA-97), LNCS 1280, 1997
- Pawlak Z, Orłowski E. Expressive power of knowledge representation. Intl. J. of Man-Machine Studies, 1984(20)
- Slowinski R, Stefanowski J. 'RoughDAS' and 'RoughClass' software implementations of the rough sets approach. In: Slowinski R ed. Intelligent Decision Support; Handbook of Applications and Advances of rough sets Theory. Kluwer Academic Pub., 1992
- Kerber R. ChiMerge: Discretization of Numeric Attributes. In: Proceedings of 10th National Conf. on Artificial Intelligence (AAAI 92). MIT Press, 1992
- 史忠植. 高级人工智能. 科学出版社, 1997

(上接第 58 页)