水产品安全信息系统中属性离散化方法研究

鄂 旭[1],杨 健[2],王欣铨[3],刘忠杰[1],孙德才[1]
(1. 渤海大学 信息科学与技术学院,辽宁 锦州 121001;
2. 中航飞机起落架有限责任公司,湖南 长沙 410200;
3. 渤海大学 艺术设计学院,辽宁 锦州 121001)

摘 要:连续属性离散化作为水产品安全信息系统中进行智能化数据处理的一个重要研究内容,已然成为水产品安全信息化研究领域的一个热点和难点。文中利用基于粗糙集理论相对熵的连续属性离散化方法来解决这个问题。此方法选用候选区间的类信息熵作为离散门限值边界,并且通过考察每个属性值的分类能力,合并离散区间,去掉冗余断点,确定关键离散属性值,最终在水产品安全信息系统中实现连续属性离散化。实例分析表明算法是有效可行的。

关键词:粗糙集;离散化;食品安全;信息系统
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2014)07-0178-03
doi:10.3969/j.issn.1673-629X.2014.07.044

Research on Discretization Method in Aquatic Product Safety Information System

E Xu[1], YANG Jian[2], WANG Xin-quan[3], LIU Zhong-jie[1], SUN De-cai[1]
(1. College of Information Science & Technology, Bohai University, Jinzhou 121001, China;
2. AVIC Landing Gear Manufacturing Corp., Changsha 410200, China;
3. College of Art Design, Bohai University, Jinzhou, 121001, China)

Abstract: Continuous attributes discretization is an important work for intelligent analysis in an aquatic product safety information system based on rough sets theory, and it is a hot and hard problem in research fields. To deal with it, a new discretization method was proposed based on relative entropy in rough set. The method took the candidate interval classification entropy as the discretization threshold values, and determined the key discretization values through merging the interval values and deleting the redundant values in order to realize the continuous attributes discretization in aquatic product safety information system. Experimental results indicate the algorithm is effective and feasible.

Key words: rough set; discretization; food safety; information system

0 引 言

目前,日益频发的严重食品安全事件已经引起了全世界各国人民的高度关注,食品安全也成为世界各国政府共同面临的一个首要难题。各个国家都在借助信息技术对食品的生产、运输、加工等各个环节进行全方位的监督、管理和控制,利用数据库保存了大量的数据[1]。这些数据都蕴含着大量宝贵有用的信息,与食品安全监管、评价、预警等紧密相关。因此,需要采用数据挖掘等技术,对食品安全信息进行监督、管理和分析。

连续属性离散化作为数据挖掘和机器学习中预处理的重要步骤,对数据挖掘和机器学习的最终效果具有直接关系[2-6]。但在数据挖掘和机器学习领域,许多算法只能处理一些离散化的数据。然而,在实际的数据库中,离散化的数据较少,连续属性较多。为了能够将好的数据样本从这些含有连续属性的数据库中提取出来,获得简洁有效的规则,就要将连续属性进行离散化[7-11]。离散化结果将使系统对存储空间的实际需

收稿日期:2013-06-24 **修回日期**:2013-10-15 **网络出版时间**:2014-02-24
基金项目:中国博士后基金项目(2012M520158);辽宁省百千万人才基金择优资助项目(2012921058);辽宁省教育科研项目(L2012397,L2012396,L2012400);辽宁省社科联 2014 年度辽宁经济社会发展立项课题(2014LSLKTDGLX-02)
作者简介:鄂 旭(1971-),男,教授,博士,硕士生导师,研究方向为数据挖掘与食品安全物联网。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0857.001.html

求减小,使后继数据挖掘和机器学习算法的运行速度加快,后继算法的空间开销减小,分类精度提高。连续属性的最优离散化被看作为一个 NP 完全问题,不同的离散化方法,往往可使结果存在差异。目前,离散化问题已成为学术界一项重要的研究内容[12-14]。

1　相关重要概念与定理

该离散方法涉及的主要概念和定理如下:

定义 1:设 $K=(U,R)$ 是一近似空间,R 在 U 上的划分为 $U/\mathrm{IND}(R)=\{R_1,R_2,\cdots,R_n\}$,知识(属性集合)$R$ 的信息量(也称为信息熵)定义为:

$$E(R)\sum_{i=1}^{n}\frac{|R_i|}{|U|}\frac{|R_i^c|}{|U|}=\sum_{i=1}^{n}\frac{|R_i|}{|U|}(1-\frac{|R_i|}{|U|})$$

其中,$R_i^c=U-R_i$ 表示论域上存在等价类的可能性,$|R_i^c|/|U|$ 为论域 U 上存在 R_i 的概率,也即不属于 R_i 的概率。

定义 2:假定用 S 代表样本集合、C 代表属性、T 代表区间边界。其中 S 被 T 划分为两个区间 S_1 和 S_2,$\mathrm{Ent}(S_1)$ 和 $\mathrm{Ent}(S_2)$ 分别代表其所对应区间的类信息熵,则 T 所产生的类信息熵可以表示为:$E(C,T,S)=\frac{S_1}{S}\mathrm{Ent}(S_1)+\frac{S_2}{S}\mathrm{Ent}(S_2)$。

对于给定的属性 C,所有的候选断点中最好用最小的划分点表示,记为 T,使其成为一个离散化的划分点。这样就可以二分属性 C 的区间,使样本集合被划分成两个子集合,其对应的类信息熵分别为和;如果此时对应的类信息熵最小,则继续合并;否则继续划分,即划分它们中类信息熵大的部分。利用这个方法不断地进行递归,达到满足条件为止,这样 C 的离散化区间即可得到。

定义 3:设 U 为论域,$T_1=(U,P)$ 和 $T_2=(U,Q)$ 是关于 U 的两个知识库 $U/\mathrm{IND}(P)=\{X_1,X_2,\cdots,X_n\}$,$U/\mathrm{IND}(Q)=\{Y_1,Y_2,\cdots,Y_n\}$,知识 Q 相对于 P 的相对熵 $E(Q\mid P)$ 可定义为:

$$E(Q\mid P)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|X_i\cap Y_j|}{|U|}\frac{|Y_j^c-X_i^c|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|X_i\cap Y_j|}{|U|}\frac{(|X_i-|X_i\cap Y_j||)}{|U|}$$

定理 1:设决策表为 $S=(U,A,V,f)$,其中,论域 U 是一个非空有限对象集合,A 是对象的属性集合,分为条件属性集 C 和决策属性集 D 两个不相交的子集。$\forall B\subseteq C$,令 $U'=U-\mathrm{POS}_B^U(D)$ 为粗糙边界,如果 $U'/\mathrm{IND}(B)\cup\{c\}=\{m_1,m_2,\cdots,m_p\}$,$U'/\mathrm{IND}(D)=\{n_1,n_2,\cdots,n_q\}$,则 $\forall c\in C,\mathrm{POS}_{B\cup\{c\}}^{U'}(D)=\bigcup_{i=1}^{m}\mathrm{POS}_{B\cup\{c\}}^{U'}(D)_{n_i}$。

证明:$\mathrm{POS}_{B\cup\{c\}}^{U'}(D)=\bigcup_{j=1}^{m}\underline{B\cup\{c\}}(X_j)=$

$\bigcup_{j=1}^{m}\{Y_i\in U'/B\cup\{c\}:Y_i\subseteq X_j\}=$

$\bigcup_{j=1}^{m}\{Y_i\in\{Y_1\}\cup\{Y_2\}\cup\cdots\cup\{Y_n\}:Y_i\subseteq X_j\}=$

$\bigcup_{j=1}^{m}\{Y_i\in\{Y_i\}:Y_i\subseteq X_j\}=$

$\bigcup_{j=1}^{m}\mathrm{POS}_{B\cup\{c\}}^{U'}(D)_{n_i}$

定理 2:设 $\forall B\subseteq C,\forall c\in C$ 且 $\notin B$,$U'=U-\mathrm{POS}_B^U(D)$ 为粗糙边界,则有:

$|\mathrm{POS}_{B\cup\{c\}}^{U'}(D)|=$

$|\mathrm{POS}_{B\cup\{c\}}^{U'}(D)|-|\mathrm{POS}_B^U(D)|$

证明:对 $\forall c\in C$ 且 $\notin B$ 有两种情况:

若 c 为冗余属性,则 $|\mathrm{POS}_{B\cup\{c\}}^{U'}(D)|=|\mathrm{POS}_B^U(D)|=0$,$|\mathrm{POS}_{B\cup\{c\}}^{U'}(D)|=|\mathrm{POS}_B^U(D)|-|\mathrm{POS}_B^U(D)|=0$;

若 c 为重要属性,因为 $U'=U-\mathrm{POS}_B(D)$,所以 $U=U'+\mathrm{POS}_B(D)$,所以有:$\mathrm{POS}_{B\cup\{c\}}^{U'}(D)=\mathrm{POS}_B^U(D)\cup\mathrm{POS}_{B\cup\{c\}}^{U'}(D)$,即 $|\mathrm{POS}_{B\cup\{c\}}^{U'}(D)|=|\mathrm{POS}_{B\cup\{c\}}^{U'}(D)|-|\mathrm{POS}_B^U(D)|$。

定理 3:设 $\forall B\subseteq C,\forall a,b\in C$ 且 $\notin B$,如果属性 a 比属性 b 重要,则有 $\mathrm{sig}(a)-\mathrm{sig}(b)>0$。

证明:如果属性 a 比属性 b 重要,由定理 2 可知属性 a 比属性 b 在粗糙边界的分类能力强,即 $\mathrm{POS}_{B\cup\{a\}}(D)>\mathrm{POS}_{B\cup\{b\}}(D)$,所以有:

$\mathrm{sig}(a)-\mathrm{sig}(b)=|\mathrm{POS}_{B\cup\{a\}}(D)-\mathrm{POS}_B(D)|/|U-\mathrm{POS}_B(D)|-|\mathrm{POS}_{B\cup\{b\}}(D)-\mathrm{POS}_B(D)|/|U-\mathrm{POS}_B(D)|>0$

定理证毕。

2　该离散化方法核心思想

设 $S=(U,R,V,f)$ 是一个决策表,其中论域 $U=\{x_1,x_2,\cdots,x_n\}$ 为有限个对象的集合,$R=C\cup\{d\}$,$C=\{a_1,a_2,\cdots,a_m\}$ 为条件属性集合且 $|C|=k$,$\{d\}$ 为决策属性,现假设决策种类的个数为 $r(d)$。属性 a 的值域 $V_a=[l_a,r_a]$ 上的一个断点可记为 (a,c),其中 $a\in R$,c 为实数值,c 被称作 a 上的一个断点。在 $V_a=[l_a,r_a]$ 上的任意一个断点集合 $\{(a,c_1^a),(a,c_2^a),\cdots,(a,c_{m_a}^a)\}$ 中,定义 V_a 上的一个分类 P_a:

$P_a=\{[c_0^a,c_1^a],[c_1^a,c_2^a],\cdots,[c_{m_a}^a,c_{m_{a+1}}^a]\}$

$l_a=c_0^a<c_1^a<\cdots<c_{m_a}^a<c_{m_{a+1}}^a=r_a$

$V_a=[c_0^a,c_1^a)\cup[c_1^a,c_2^a)\cup\cdots\cup[c_{m_a}^a,c_{m_{a+1}}^a]$

因此,任意的 $P=\cup P_a$ 定义了一个新的决策表 $S^p=(U,C^p\cup\{d\})$,$C^p=\{a^p:a^p(x)=i\Leftrightarrow a(x)\in[c_i^a,$

$c_{i+1}^{a}]\}$，对于 $x \in U$，$i \in \{0,\cdots,m_n\}$，新的决策系统将取代原来的决策系统。

该离散化方法离散的门限值边界是通过候选区间的类信息熵来选择的，如果存在一个最小的区间边界使熵函数最小，那么该方法就可以被递归地使用在边界所划分而产生的两个区间里。本质是如何根据选取的断点对条件属性构成的空间进行划分，将 m（m 为条件属性的个数）维空间划分为有限个具有相同决策值的区域。

3 实例分析

为了更好地描述该离散方法，设表 1 为原始水产品安全信息表，其中决策属性 $A=\{a,b\}$，条件属性 $D=\{d\}$。

表 1 原始水产品安全信息表

U	a	b	d
1	0.4	2	1
2	0.8	0.4	0
3	1	2	0
4	1.2	1	1
5	1.2	3	0
6	1	1	1
7	2	3	1
8	4	3	1

显然有：

$U/\{d\} = \{\{1,4,6,7,8\},\{2,3,5\}\} = \{Y_1,Y_2\}$；

$U/\{a\} = \{\{1\},\{2\},\{3,6\},\{4,5\},\{7\},\{8\}\} = \{X_1,X_2,X_3,X_4,X_5,X_6\}$；

$U/\{b\} = \{\{1,3\},\{2\},\{4,6\},\{5,7,8\}\} = \{Z_1,Z_2,Z_3,Z_4\}$。

计算属性重要性：

$$E(D \mid a) = \frac{1}{8*8}(0+0+2+2+0+0) = \frac{1}{16};$$

$$E(D \mid b) = \frac{1}{8*8}(2+0+0+4) = \frac{3}{32}。$$

可以看出属性 $\mathrm{SGF}(a,D) = E(D) - E(D \mid a) > E(D) - E(D \mid b) = \mathrm{SGF}(b,D)$。

对决策表 1，其候选断点集为 $C_a = \{0.6,0.9,1.1,1.5,3\}$，$C_b = \{0.7,1.2,1.5,2,2.5\}$。

因此，应该从属性 b 开始考虑，首先考虑 0.7，它有相邻属性 0.4 和 1，把 0.4 改为 1 后不会引起决策表的冲突，因此断点 0.7 是多余的；再考虑 2，它的相邻属性为 1 和 3，若将 1 改为 3 则会引起实例 4 和实例 5 相冲突，故断点 2 是保持决策表不可分辨关系所必须具有的断点值。以此类推可知断点 1.2、1.5 和 2.5 均是多余的，由此得到表 2。

同样对属性 a 的断点进行判断，得到表 3，最终得到表 4。

表 2 离散化表-1

U	a	b	d
1	0.4	3	1
2	0.8	1	0
3	1	3	0
4	1.2	1	1
5	1.2	3	0
6	1	1	1
7	2	3	1
8	4	3	1

表 3 离散化表-2

U	a	b	d
1	1	3	1
2	1	1	0
3	1	3	0
4	1.2	1	1
5	1.2	3	0
6	1.2	1	1
7	4	3	1
8	4	3	1

表 4 最终离散化表

U	a	b	d
1	0	1	1
2	0	0	0
3	0	1	0
4	1	0	1
5	1	1	0
6	1	0	1
7	2	1	1
8	2	1	1

4 结束语

该连续属性离散化方法的离散化过程将属性值的具体情况考虑在内，克服了粗糙集中离散化方法存在的缺陷，并且能够在进行离散化的过程中删除掉无用的属性，维持了离散前后系统信息的一致性。

参考文献：

[1] 邓聪文，朱雪冬，王俊能. 食品安全评价及其方法简述[J]. 食品安全，2009(6)：8-10.

（下转第 184 页）

相同的 SVM 分类器。另外,由于文中算法运用将最容易分离的类最先分割出来的策略,从而获得的分类精度要比其他分类算法要高,同时训练时间也比 OVR 算法短,表明了该算法是可行的。

4 结束语

文中在研究了现有的 SVM 多分类的算法的基础上,提出了基于相对距离的 BT-SVM 多分类算法,用类中心的相对距离来衡量两个类之间的差异,在构造出二叉树结构过程中,最容易分离的类最先被分割出来。实验结果表明,该算法与 OVR、OVO 方法相比,能获得较高的分类精度和较低的分类时间。虽然基于相对距离的 BT-SVM 多分类算法在多分类问题上取得了较好的效果,但是分类的时间仍较长,所以该课题接下来将致力于缩短分类的时间。

参考文献:

[1] Vapnik V N. The nature of statistical learning theory[M]. New York:Springer-Verlag,1995.

[2] Byun H,Lee S W. Applications of support vector machines for pattern recognition: a survey[J]. LNCS, 2002, 2388: 213-236.

[3] 朱凤明,樊明龙. 混沌粒子群算法对支持向量机模型参数的优化[J]. 计算机仿真,2010,27(11):183-186.

[4] 朱志慧,李 雷,种冬梅. 改进的 BT-SVM 应用于电力系统 SSA[J]. 计算机技术与发展,2012,22(9):157-160.

[5] Weston J, Watkins C. Multi-class support vector machines [C]//Proceedings of ESANN. [s. l.]:[s. n.],1999.

[6] Rifkin R M,Klautau A. In defense of one-vs-all classification [J]. Journal of Machine Learning Research, 2004, 5: 101-141.

[7] 赵有星,李 琳. 基于支持向量机的多类分类算法研究[J]. 科技信息,2007(29):129-130.

[8] Platt J C, Cristianini N, Shawe-taylor J. Large margin DAGs for multiclass classification[C]//Advances in neural information processing systems. Cambridge, MA: MIT Press, 2000: 547-553.

[9] 王 艳,陈欢欢,沈 毅. 有向无环图的多类支持向量机分类算法[J]. 电机与控制学报,2011,15(4):85-89.

[10] Cheong S. Support vector machines with binary tree architecture for multi-class classification[J]. Neural Information Processing-Letters and Reviews, 2004, 2(3): 47-51.

[11] 范柏超,王建宇,薄煜明. 结合特征选择的二叉树 SVM 多分类算法[J]. 计算机工程与设计,2010,31(12):2823-2825.

[12] Hsu Chih-Wei, Lin Chih-Jen. A comparison of methods for multi-class support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.

[13] Takahashi F, Abe S. Decision-tree-based multiclass support vector machines[C]//Proceedings of the 9th international conference on neural information processing. [s. l.]: IEEE, 2002: 1418-1422.

[14] 唐发明,王仲东,陈绵云. 支持向量机多类分类算法研究[J]. 控制与决策,2005,20(7):746-749.

[15] 刘 健,刘 忠,熊 鹰. 改进的二叉树支持向量机多类分类算法研究[J]. 计算机工程与应用,2010,46(33):117-120.

[16] Tax D M J, Duin R P W. Support vector domain description [J]. Pattern Recognition Letters, 1999, 20: 1191-1199.

(上接第 180 页)

[2] Pawlak Z. Roughset[J]. International Journal of Computer and Information Science, 1982, 11(5): 341-356.

[3] 武 森,高学东,Bastian M. 数据仓库与数据挖掘[M]. 北京:冶金工业出版社,2003.

[4] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2003.

[5] 苗夺谦. Rough Set 理论中连续属性的离散化方法[J]. 自动化学报,2001,27(3):296-302.

[6] Nguyen H S, Skowron A. Boolean reasoning for feature extraction problems[C]//Proc of 10th international symposium on foundations of intelligent systems. New York: Springer-Verlag, 1997: 117-126.

[7] 鄂 旭,高学东,谭文东,等. 基于超立方体与信息熵的离散化方法[J]. 北京科技大学学报,2005,27(6):760-763.

[8] 孟庆生. 信息论[M]. 西安:西安交通大学出版社,1986.

[9] Kryszkiewicz M. Rules in incomplete information system[J]. Information Science, 1999, 113(3-4): 271-292.

[10] 李仁璞,黄 道. 基于 RS 理论的不完备信息系统处理方法[J]. 华东理工大学学报(自然科学版),2005,31(2):227-231.

[11] 邓耀进,李仁发. 一种粗糙集理论中量化容差关系的改进[J]. 计算机工程与科学,2009,31(10):105-107.

[12] 杨霁琳,秦克云,裴 峥. 不完备信息系统中的不可区分关系[J]. 计算机工程,2010,36(13):4-6.

[13] 鄂 旭,高学东,喻 斌. 基于扫描向量的属性约简方法[J]. 北京科技大学学报,2006,28(6):604-608.

[14] E Xu, Yang Yuqiang, Ren Yongchang. A new method of attribute reduction based on information quantity in an incomplete system[J]. Journal of Software, 2012, 7(8): 1881-1888.

水产品安全信息系统中属性离散化方法研究

作者： 鄂旭， 杨健， 王欣铨， 刘忠杰， 孙德才， E Xu， YANG Jian， WANG Xin-quan， LIU Zhong-jie， SUN De-cai

作者单位： 鄂旭,刘忠杰,孙德才,E Xu,LIU Zhong-jie,SUN De-cai(渤海大学 信息科学与技术学院,辽宁 锦州,121001)， 杨健,YANG Jian(中航飞机起落架有限责任公司,湖南 长沙,410200)， 王欣铨,WANG Xin-quan(渤海大学 艺术设计学院,辽宁 锦州,121001)

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(7)

参考文献(14条)

1. 邓聪文;朱雪冬;王俊能 食品安全评价及其方法简述 2009(06)

2. Pawlak Z Roughset 1982(05)

3. 武森;高学东;Bastian M 数据仓库与数据挖掘 2003

4. 王国胤 Rough集理论与知识获取 2003

5. 苗夺谦 Rough Set理论中连续属性的离散化方法 2001(03)

6. Nguyen H S;Skowron A Boolean reasoning for feature extrac-tion problems 1997

7. 鄂旭;高学东;谭文东 基于超立方体与信息熵的离散化方法 2005(06)

8. 孟庆生 信息论 1986

9. Kryszkiewicz M Rules in incomplete information system 1999(3-4)

10. 李仁璞;黄道 基于RS理论的不完备信息系统处理方法 2005(02)

11. 邓耀进;李仁发 一种粗糙集理论中量化容差关系的改进 2009(10)

12. 杨霁琳;秦克云;裴峥 不完备信息系统中的不可区分关系 2010(13)

13. 鄂旭;高学东;喻斌 基于扫描向量的属性约简方法 2006(06)

14. E Xu;Yang Yuqiang;Ren Yongchang A new method of attrib-ute reduction based on information quantity in an incomplete system 2012(08)