知识挖掘与它的链式特征

孟令存[1],刘月兰[2],郝秀梅[3]
MENG Ling-cun[1],LIU Yue-lan[2],HAO Xiu-mei[3]

1.济宁职业技术学院,山东 济宁 232037
2.山东省青年管理干部学院 国际贸易系,济南 250014
3.山东财政学院 统计与数理学院,济南 250014
1.Jining Vocational and Technical College,Jining,Shandong 232037,China
2.Department of International Trade,Shandong Youth Administrative College,Ji'nan 250014,China
3.Department of Statistics and Mathematics,Shandong Finance Institute,Ji'nan 250014,China
E-mail:mlcwfm@126.com

MENG Ling-cun,LIU Yue-lan,HAO Xiu-mei.Knowledge mining and its chain type properties.Computer Engineering and Applications,2007,43(19):42-44.

Abstract: This paper presents the concepts of $f,\bar{f}$ knowledge and mining degree by using S-rough set and its attribute transfer, and discusses the relation between attribute transfer and knowledge mining data property.This paper gives some knowledge chain theorems and minimum,maximum mining degree theorem of f, $\bar{f}$.Finally,this paper presents an analysis on $\bar{f}$ knowledge mining example.

Key words: one direction S-rough set;attribute transfer;mining degree;mining dependence theorem

摘 要:利用S-粗集与它的属性迁移,提出$f,\bar{f}$知识、挖掘度概念,讨论了属性迁移与知识挖掘的数量关系;给出了$f,\bar{f}$知识链式定理和$f,\bar{f}$知识最小、最大挖掘度定理。最后,给出了$\bar{f}$知识挖掘的实例分析。

关键词:单向S-粗集;属性迁移;挖掘度;挖掘定理
文章编号:1002-8331(2007)19-0042-03 **文献标识码**:A **中图分类号**:TP311

1 引言

1982年波兰数学家Z.Pawlak教授提出粗集(Rough Sets)[1,2],给出了粗集的一般性研究。粗集在静态数据挖掘与静态知识发现中得到了应用,但是,由于系统受到外界因素的攻击,$X\subseteq U$经常变动,也就是说X具有动态特性,这就使得Z.Pawlak粗集研究动态$X\subseteq U$遇到了困难,有一定的局限性。2002年史开泉教授提出了S-粗集(Singular Rough Sets)[3],给出了S-粗集一般讨论,粗集理论获得了广泛的应用[4-11]。本文给出了S-粗集在动态数据挖掘与知识发现的理论研究及具体应用, 提出了$f,\bar{f}$迁移知识、知识挖掘度概念,研究了与知识挖掘有关的各种因素及数量特征并给出相应链式定理。

为了使本文符号简化,又不致引起误解,这里约定:U是有限元素论域;V是U对应的有限属性论域,$F=\{f_1,f_2,\cdots,f_n\}$,$\bar{F}=\{\bar{f}_1,\bar{f}_2,\cdots,\bar{f}_n\}$是$U$上的元素迁移族,$X\subseteq U$,$\alpha$是属性集;$[x]_\alpha$是具有$\alpha$的等价类、知识、等价关系,属性不加区分,直接使用。

2 $(f,\bar{f})$知识挖掘与其数量特征

定义1 设$\alpha=\{\alpha_1,\alpha_2,\cdots,\alpha_\lambda\}$是$V$上的属性集,$f\in F,\bar{f}\in\bar{F}$是$V$上的元素迁移,分别称$\alpha^f$、$\alpha^{\bar{f}}$是$\alpha$的属性补充集、属性删除集,如果

$$\alpha^f=\alpha\cup\{f_i(\beta_i)=\alpha_i'\},\beta_i\in V,\beta_i\notin\alpha,f(\beta_i)=\alpha_i'\in\alpha$$

$$\alpha^{\bar{f}}=\alpha\backslash\{f(\alpha_i)=\beta_i\},\alpha_i\in\alpha,f(\alpha_i)=\beta_i\notin\alpha\quad i=1,2,\cdots,m$$

定义2 称$[x]_\alpha^f$是知识$[x]_\alpha$的f知识,如果$[x]_\alpha^f$的属性集α^f与$[x]_\alpha$的属性集α满足$card(\alpha)\leqslant card(\alpha^f)$,其中$card(\alpha)$、$card(\alpha^f)$分别是$\alpha$、$\alpha^f$的基数。

定义3 称$[x]_\alpha^{\bar{f}}$是知识$[x]_\alpha$的$\bar{f}$知识,如果$[x]_\alpha^{\bar{f}}$的属性集$\alpha^{\bar{f}}$与$[x]_\alpha$的属性集α满足

$$card(\alpha^{\bar{f}})\leqslant card(\alpha)$$

基金项目:山东省自然科学基金(the Natural Science Foundation of Shandong Province of China under Grant No.Y2004A04);山东省科学发展计划项目(the Science Development Plan of Shandong Province of China under Grant No.B2006053)。

作者简介:孟令存(1965-),男,讲师,研究方向:数据挖掘,粗系统理论与应用;刘月兰(1963-),女,副教授,研究方向:粗系统理论与应用;郝秀梅(1965-),女,教授,研究方向:粗糙集。

这里 $card(\alpha^{\bar{f}})$ 是 $\alpha^{\bar{f}}$ 的基数。

定义 4 称 $[x]_{(\alpha\cup\{f(\beta_1)\})}$ 是 $[x]_\alpha$ 的 1 阶 f 知识,如果存在 $\beta_1\in\alpha$, $f(\beta_1)=\alpha_1'\in\alpha$;称 $[x]_{(\alpha\cup\{f(\beta_1),f(\beta_2),\cdots,f(\beta_k)\})}$ 是 $[x]_\alpha$ 的 k 阶 f 知识,记为 $[x]^k_{(\alpha,f)}$;称 $[x]_{(\alpha\setminus\{\bar{f}(\alpha_i)\})}$ 是 $[x]_\alpha$ 的 1 阶 $\bar{f}$ 知识,如果存在 $\alpha_i\in\alpha$, $\bar{f}(\alpha_i)=\beta_i\in\alpha$;称 $[x]_{(\alpha\setminus\{\bar{f}(\alpha_1),\bar{f}(\alpha_2),\cdots,\bar{f}(\alpha_k)\})}$ 是 $[x]_\alpha$ 的 k 阶 $\bar{f}$ 知识,记为 $[x]^k_{(\alpha,\bar{f})}$(k 为正整数)。

定义 5 称 $DMD([x]^k_{(\alpha,f)})$ 是 k 阶 f 知识 $[x]^k_{(\alpha,f)}$ 关于知识 $[x]_\alpha$ 的 f 挖掘度(mining degree),如果

$$DMD([x]^k_{(\alpha,f)})=\frac{GRD([x]^k_{(\alpha,f)})}{GRD([x]_\alpha)}$$

称 $DMD([x]^k_{(\alpha,\bar{f})})$ 是 k 阶 $\bar{f}$ 知识 $[x]^k_{(\alpha,\bar{f})}$ 关于知识 $[x]_\alpha$ 的 $\bar{f}$ 挖掘度(mining degree),如果

$$DMD([x]^k_{(\alpha,\bar{f})})=\frac{GRD([x]_\alpha)}{GRD([x]^k_{(\alpha,\bar{f})})}$$

命题 1 f 知识 $[x]^f_\alpha$ 存在于知识 $[x]_\alpha$ 中。

命题 2 $\cdots\subseteq[x]^k_{(\alpha,f)}\subseteq[x]^{k-1}_{(\alpha,f)}\subseteq\cdots\subseteq[x]^2_{(\alpha,f)}\subseteq[x]^1_{(\alpha,f)}\subseteq[x]_\alpha\subseteq X\subset U$。

命题 3 $[x]_\alpha$ 存在于所有 $\bar{f}$ 知识 $[x]^{\bar{f}}_\alpha$ 中。

命题 4 $[x]_\alpha\subseteq[x]^1_{(\alpha,\bar{f})}\subseteq[x]^2_{(\alpha,\bar{f})}\subseteq\cdots\subseteq[x]^{k-1}_{(\alpha,\bar{f})}\subseteq[x]^k_{(\alpha,\bar{f})}\subseteq\cdots\subseteq U$。

命题 5 任意 $[x]^k_{(\alpha,f)}\cap[x]^j_{(\alpha,f)}\neq\Phi,(k\neq j)$。

定理 1 f 知识挖掘基数链定理:设 $[x]^k_{(\alpha,f)}$ 是 $[x]_\alpha$ 的 k 阶 f 知识,$k=k_1k_2\cdots k_t$,若 $k_1\leqslant k_2\leqslant\cdots\leqslant k_{t-1}\leqslant k_t$,则

$$card([x]^{k_t}_{(\alpha,f)})\leqslant card([x]^{k_{t-1}}_{(\alpha,f)})\leqslant\cdots\leqslant card([x]^{k_1}_{(\alpha,f)})$$

证明 因为 $k_1\leqslant k_2\leqslant\cdots\leqslant k_{t-1}\leqslant k_t$,则有

$$(\alpha\cup\{f(\beta_1),\cdots,f(\beta_{k_1})\})\subseteq(\alpha\cup\{f(\beta_1),\cdots,f(\beta_{k_2})\})\subseteq\cdots\subseteq(\alpha\cup\{f(\beta_1),\cdots,f(\beta_{k_t})\})$$

因此

$$[x]^{k_t}_{(\alpha,f)}\subseteq[x]^{k_{t-1}}_{(\alpha,f)}\subseteq\cdots\subseteq[x]^{k_1}_{(\alpha,f)}$$

$$card([x]^{k_t}_{(\alpha,f)})\leqslant card([x]^{k_{t-1}}_{(\alpha,f)})\leqslant\cdots\leqslant card([x]^{k_1}_{(\alpha,f)})$$

定理 2 f 知识挖掘粒度链定理:设 $[x]^k_{(\alpha,f)}$ 是 $[x]_\alpha$ 的 k 阶 f 知识,$k=k_1k_2\cdots k_t$,若 $k_1\leqslant k_2\leqslant\cdots\leqslant k_{t-1}\leqslant k_t$,则

$$GRD([x]^{k_t}_{(\alpha,f)})\leqslant GRD([x]^{k_{t-1}}_{(\alpha,f)})\leqslant\cdots\leqslant GRD([x]^{k_1}_{(\alpha,f)})$$

推论 1 $\bar{f}$ 知识挖掘的粒度链定理:设 $[x]^k_{(\alpha,\bar{f})}$ 是 $[x]_\alpha$ 的 k 阶 $\bar{f}$ 知识,$k=k_1k_2\cdots k_t$,若 $k_1\leqslant k_2\leqslant\cdots\leqslant k_{t-1}\leqslant k_t$,则

$$GRD([x]^{k_t}_{(\alpha,\bar{f})})\geqslant GRD([x]^{k_{t-1}}_{(\alpha,\bar{f})})\geqslant\cdots\geqslant GRD([x]^{k_1}_{(\alpha,\bar{f})})$$

由上述定理、推论及知识过滤度定义很容易得到以下两个定理:

定理 3 f 知识挖掘过滤度链定理:设 $[x]^k_{(\alpha,f)}$ 是 $[x]_\alpha$ 的 k 阶 f 知识,$k=k_1k_2\cdots k_t$,若 $k_1\leqslant k_2\leqslant\cdots\leqslant k_{t-1}\leqslant k_t$,则

$$FID([x]^{k_t}_{(\alpha,f)})\geqslant FID([x]^{k_{t-1}}_{(\alpha,f)})\geqslant\cdots\geqslant FID([x]^{k_1}_{(\alpha,f)})$$

推论 2 $\bar{f}$ 知识挖掘过滤度链定理:设 $[x]^k_{(\alpha,\bar{f})}$ 是 $[x]_\alpha$ 的 k 阶 $\bar{f}$ 知识,$k=k_1k_2\cdots k_t$,若 $k_1\leqslant k_2\leqslant\cdots\leqslant k_{t-1}\leqslant k_t$,则

$$FID([x]^{k_t}_{(\alpha,\bar{f})})\leqslant FID([x]^{k_{t-1}}_{(\alpha,\bar{f})})\leqslant\cdots\leqslant FID([x]^{k_1}_{(\alpha,\bar{f})})$$

定理 4 f 知识挖掘度链定理:设 $[x]^k_{(\alpha,f)}$ 是 $[x]_\alpha$ 的 k 阶 f 知识,$k=k_1k_2\cdots k_t$,若 $k_1\leqslant k_2\leqslant\cdots\leqslant k_{t-1}\leqslant k_t$,则

$$DMD([x]^{k_t}_{(\alpha,f)})\leqslant DMD([x]^{k_{t-1}}_{(\alpha,f)})\leqslant\cdots\leqslant DMD([x]^{k_1}_{(\alpha,f)})$$

证明 因为 $k_1\leqslant k_2\leqslant\cdots\leqslant k_t$,所以 $[x]^{k_t}_{(\alpha,f)}\subseteq[x]^{k_{t-1}}_{(\alpha,f)}\subseteq\cdots\subseteq[x]^{k_1}_{(\alpha,f)}$,从而 $GRD([x]^{k_t}_{(\alpha,f)})\leqslant GRD([x]^{k_{t-1}}_{(\alpha,f)})\leqslant\cdots\leqslant GRD([x]^{k_1}_{(\alpha,f)})$,$\frac{GRD([x]^{k_t}_{(\alpha,f)})}{GRD([x]_\alpha)}\leqslant\frac{GRD([x]^{k_{t-1}}_{(\alpha,f)})}{GRD([x]_\alpha)}\leqslant\cdots\leqslant\frac{GRD([x]^{k_1}_{(\alpha,f)})}{GRD([x]_\alpha)}$,所以

$$DMD([x]^{k_t}_{(\alpha,f)})\leqslant DMD([x]^{k_{t-1}}_{(\alpha,f)})\leqslant\cdots\leqslant DMD([x]^{k_1}_{(\alpha,f)})$$

由推论 1 及定义 5 得到:

推论 3 $\bar{f}$ 知识挖掘度链定理:设 $[x]^k_{(\alpha,\bar{f})}$ 是 $[x]_\alpha$ 的 k 阶 $\bar{f}$ 知识,$k=k_1k_2\cdots k_t$,若 $k_1\leqslant k_2\leqslant\cdots\leqslant k_{t-1}\leqslant k_t$,则

$$DMD([x]^{k_t}_{(\alpha,\bar{f})})\geqslant DMD([x]^{k_{t-1}}_{(\alpha,\bar{f})})\geqslant\cdots\geqslant DMD([x]^{k_1}_{(\alpha,\bar{f})})$$

定理 5 f 最小挖掘度定理:设 $[x]^k_{(\alpha,f)}$ 是 $[x]_\alpha$ 的 k 阶 f 知识,若 $[x]^k_{(\alpha,f)}$ 的属性集 $(\alpha\cup\alpha^f)$ 满足

$$card(\alpha\cup\alpha^f)=\lambda+k$$

则 $[x]^k_{(\alpha,f)}$ 挖掘度 $DMD([x]^k_{(\alpha,f)})$ 最小,而且

$$DMD([x]^k_{(\alpha,f)})=\min$$

证明 因为 $card(\alpha\cup\alpha^f)=\lambda+k$,则 k 阶 f 知识 $[x]^k_{(\alpha,f)}$ 挖掘阶 $k_{\max}=\max\limits_{i=1}(k_i)$,利用定理 4 得:

$$DMD([x]^{k_t}_{(\alpha,f)})\leqslant DMD([x]^{k_{t-1}}_{(\alpha,f)})\leqslant\cdots\leqslant DMD([x]^{k_1}_{(\alpha,f)})$$

$$DMD([x]^{k_t}_{(\alpha,f)})=\min$$

定理 6 $\bar{f}$ 知识最大挖掘度定理:设 $[x]^k_{(\alpha,\bar{f})}$ 是 $[x]_\alpha$ 的 k 阶 $\bar{f}$ 知识,若 $[x]^k_{(\alpha,\bar{f})}$ 的属性集 $(\alpha\setminus\alpha^{\bar{f}})$ 满足

$$card(\alpha\setminus\alpha^{\bar{f}})=1$$

则 $[x]^k_{(\alpha,\bar{f})}$ 挖掘度 $DMD([x]^k_{(\alpha,\bar{f})})$ 最大,而且

$$DMD([x]^k_{(\alpha,\bar{f})})=\max$$

证明 因为 $card(\alpha\setminus\alpha^{\bar{f}})=1$,则 k 阶 $\bar{f}$ 知识 $[x]^k_{(\alpha,\bar{f})}$ 挖掘阶 $k_{\max}=\max\limits_{i=1}(k_i)$ 最大,由推论 3 得到

$$DMD([x]^{k_t}_{(\alpha,\bar{f})})\geqslant DMD([x]^{k_{t-1}}_{(\alpha,\bar{f})})\geqslant\cdots\geqslant DMD([x]^{k_1}_{(\alpha,\bar{f})})$$

所以

$$DMD([x]^k_{(\alpha,\bar{f})})=\max$$

知识粒度与知识挖掘度关系原理:知识 $[x]_\alpha$ 的粒度越大,f 知识越多,挖掘度就越大,反之亦然。同样,知识 $[x]_\alpha$ 的粒度越小,外挖的知识越多,外挖掘度越大,反之亦然。

3 知识挖掘的应用

为了简单,下面只给出 $\bar{f}$ 知识挖掘的简单应用。本章的例子取自某光纤检测系统的一个子系统。子系统在 $t_1\sim t_4$ 的输出状

态如表1。

表1中,系统在t_1,t_2,t_3,t_4的输出状态x_1,x_2,x_3,x_4的数据用"*"表示,具体的数据,略,这样不影响对系统的分析。x_1,x_2,x_3,x_4关于t_j构成知识$[x]_\alpha$,$\alpha=\{\alpha_1,\alpha_2,\alpha_3\}$,$\alpha_1,\alpha_2,\alpha_3$的属性名称,略。系统状态在$t_1$~$t_4$稳定。预警输出在稳定状态是"-"(0输出)。子系统在t_5受到属性攻击,$\alpha_3\in\alpha$,$\bar{f}(\alpha_3)=\beta_3\in\alpha$,$\alpha^{\bar{f}}=\{\alpha_1,\alpha_2\}$,子系统$t_4$的状态变成$[x]_\alpha^{\bar{f}}=\{x_1,x_2,x_3,x_4,x_0\}$,具有属性$\alpha^{\bar{f}}=\{\alpha_1,\alpha_2\}$,显然,由于属性删除,$DMD([x]^1_{(\alpha,\bar{f})})\geqslant DMD([x]_\alpha)$,知识$[x]_\alpha^{\bar{f}}$被挖掘出来。显然,$\bar{f}$知识是藏在$[x]_\alpha$之外的。当子系统受到多个属性的入侵时,可作类似的分析。

表1 子系统的t_j的工作状态

	t_1	t_2	t_3	t_4
x_1	*	*	*	*
x_2	*	*	*	*
x_3	*	*	*	*
x_4	*	*	*	*
x_0	-	-	-	-

表2 子系统的t_j的工作状态(属性删除)

	t_2	t_3	t_4	t_5
x_1	*	*	*	*
x_2	*	*	*	*
x_3	*	*	*	*
x_4	*	*	*	*
x_0	-	-	-	*

4 结束语

本文提出了f、$\bar{f}$挖掘度概念,讨论了S-粗集知识(f、$\bar{f}$知识)的数量特性,给出知识粒度、过滤度、挖掘度链定理,知识挖掘是知识发现的一个新的研究方向。$\bar{f}$知识挖掘告诫人们,系统遇到攻击之前,人们应当预先知道系统状态如何变化,制定补救的措施,避免系统紊乱,使系统损失降低最少。事实上,控制系统中的反馈就是该方法的具体体现;另外,f知识挖掘也为人们处理工程、经济管理问题,从海量数据中提出有用的知识提供了一个很好的方法。(收稿日期:2007年4月)

参考文献:

[1] Pawlak Z.Rough sets[J].International Journal of Computer and Information Science,1982(11):341-356.

[2] Pawlak Z.Rough sets-theoretical aspects of reasoning of reasoning about deta[M].Dordrecht:Kluwer Academic Publishers,1991.

[3] 史开泉,崔玉泉.S-粗集和它的一般结构[J].山东大学学报:理学版,2002(6):471-474.

[4] 张文修.粗集理论与方法[M].北京:科学出版社,1998:16-19.

[5] Shi Kai-quan.S-rough sets and its application in diagnosis recognition for disease[C]//IEEE Proceedings of the First International Conference on Machine Learning and Cybenetics,2002,4(1):50-54.

[5] 史开泉,崔玉泉.S-粗集和它的一般结构[J].山东大学学报:理学版,2002(6):471-474.

[6] Shi Kai-quan,Cui Yu-quan.F-decomposition and $\bar{F}$-reduction of S-rough sets[J].An International Journal Advances in Systems Science and Applications,2004,4:487-499.

[7] Shi Kai-quan,Cui Yu-quan.One direction S-rough decision and its decision model[C]//IEEE Proceedings of the Third International Conference on Machine Learning and Cybenetics,2004,7(3):1352-1356.

[8] Shi Kai-quan.S-rough sets and knowledge separation[J].Journal of Systems Engineering and Electronics,2005,2:401-410.

[9] Shi Kai-quan,Chang Ting-cheng.One direction S-rough sets[J].International Journal of Fuzzy Mathematics,2005,2:319-334.

[10] Shi Kai-quan.Two direction S-rough sets[J].International Journal of Fuzzy Mathematics,2005,2:335-349.

[11] Yin Shou-feng,Hu Hai-qing,Shi Kai-quan.Rough recognition of knowledge and its applications[J].An International Journal of Advances in System Sciences and Applications,2004(1):13-22.

(上接31页)

所提取的参数具有更佳的性能。

5 结语

通过以上实验表明Bark子波变换提取的语音参数和PNN神经网络相结合能够实现说话人数字语音识别,并且相对于传统方法具有明显的优势。但如果外部出现强烈噪声的情况下,如何能够提取更为纯净的语音参数来进一步提升PNN的模式识别能力,对于这方面的工作还有待于进一步研究。

(收稿日期:2007年1月)

参考文献:

[1] Specht D F.Probabilistic neural networks[J].Neural Networks,1990,3(2):109-118.

[2] Traunmullar H.Analytical expression for the tonotopic sensory scale[J].Journal of the Acoustical Society of America,1990,88:97-100.

[3] Wilson B,Finley C C,Lawson D,et al.Better speech recognition with cochlear implants[J].Nature,1991,352:236.

[4] Huang D S.Radial basis probabilistic neural networks:model and application EJ2[J].International Journal of Pattern Recognition and Artificial Intelligence,l999,13(7):1083-1101.

[5] Huang D S.The pattern recognition system theory based on the neural networks[M].Beijing:Publishing House of Electronic Industry,1996:119-137.

[6] McDermott H,Mc Kay C,Vandali A.A new portable sound processor for the University of Melbourne/Nucleus Limited multielectrode cochlear implant[J].Journal of the Acoustical Society of America,1992,91(6):3367-3371.

[7] Wallenberger E,Battmer R.Comparative speech recognition results in eight subjects using two different coding strategies with the Nucleus 22 channel cochlear implant[J].British Journal of Audiolog,1991,25:371-380.

[8] 易克初,田斌,付强.语音信号处理[M].北京:国防工业出版社,2000.

[9] 付强,易克初.语音信号的Bark子波变换及其在语音识别中的应用[J].电子学报,2000,28(10):102-105.

[10] 陶智,赵鹤鸣,龚呈卉.基于听觉掩蔽效应和Bark子波变换的语音增强[J].声学学报,2005,30(4):367-372.