基于属性重要性和样本信息熵的多连续属性离散化后处理方法

康曙光[1],裴志利[1],孔　英[2]

(1.内蒙古民族大学 计算机科学与技术学院,内蒙古 通辽　028043;2.大连医科大学,辽宁 大连　116027)

〔摘　要〕有效判别决策表中离散化后样本数据的类型对于对后继阶段的机器学习和数据挖掘过程具有非常重要的意义.本文提出了一种基于属性重要性和样本信息熵的数据类型判别方法,并利用人工改造的一部分UCI数据库进行了模拟试验,结果表明方法是有效的,识别样本数据的准确率较高、识别错误率和拒识率较低.

〔关键词〕决策表;离散化;属性重要性;样本信息熵
〔中图分类号〕TP301.6　〔文献标识码〕A　〔文章编号〕1671-0185(2009)02-0143-02

Treatment Method after Discretization of Continuous Attributes Based on Attributes Importance and Samples Entropy

KANG Shu-guang[1], PEI Zhi-li, KONG Ying[2]
(1.College of Computer Science and Technolog, Inner Mongolia University for Nationalities, Tongliao 028043,China; 2.Dalian Medical University, Dalian 116027, China)

Abstract: It has great significance to efficiently distinguish the type of the samples' data in the decision table after the discretization for the course of machine learning and data mining afterwards. This paper puts forward an annotation method of distinguishing the data type based on attributes importance and the samples entropy, and processed the simulation test using part of the UCI database which was artificially modified, it turns out the method is able to efficiently identify the data type with high accuracy, low misidentification rate and low reject rate.

Key words: Decision table; Discretization;Attributes importance; Samples entropy

0　引言

决策系统中连续属性的离散化,可以显著提高学习结果对样本的聚类能力,增强对数据噪音的鲁棒性;离散化结果将会减小系统对存储空间的实际需求,加快后继算法的运行速度;离散化过程可能够将某一连续属性的所有属性值均映射为同一结果,该属性可以被删除,系统的结构得以被简化;许多机器学习算法本身就是针对离散符号空间提出的,连续属性的离散化是应用这类算法的前提.比较有影响的是Skowron等人提出的粗糙集与布尔逻辑方法[1],该方法具有完备性,即理论上可以找出所有可能组合的离散化断点集,但是其算法复杂度是指数级的,无法在实际问题中应用,为此文献[2~6]提出了在此基础上的几种改进贪心算法,这些算法都是基于断点对实例的可分性,属于局部寻优搜索算法,而文献[7,8]则采用遗传

收稿日期:2008-07-10
基金项目:国家自然科学基金资助项目(30400162)
作者简介:康曙光(1970-),男(蒙古族),内蒙古通辽人,实验师.

算法搜索最佳离散化断点集合,属于整体搜索算法,文献[9]给出了一种将多项式超曲面与支撑向量机方法相结合的离散化方法,文献〔10〕提出了一种基于云模型的离散化方法.

上述的很多离散化算法都具有以下特征:有监督、鲁棒性、针对单一属性.这些特征容易造成一些异常数据被当作噪声数据忽略,而一些错误数据则由于没有及时清除而影响了分类结果.本文针对这一问题提出了基于属性重要度和样本信息熵的多连续属性离散化的后处理方法,通过对改造的一部分UCI数据库进行模拟试验,实现了对正常数据、噪声数据、错误数据和异常数据的高准确率识别.

1 相关概念

1.1 基于粗糙集理论属性重要性的相关概念

定义1 决策表DT.设四元组$S=<U,A,V,f>$为近似空间,其中,$U=\{x_1,x_2,\cdots,x_n\}$为非空有限论域;$A=CYD$是U的条件属性和决策属性的并集,且$C\cap D=\phi$;V是属性的值域;$f:U\times A\to V$是信息函数,指定U中每个对象x_i的属性值.R为属性或属性的集合形成的等价关系,对论域划分生成一系列的等价类.

定义2 $U/R=\{X_1,X_2,\cdots,X_n\}$,对于$Y\subseteq U$,定义$Y$关于关系$R$的下近似集、上近似集为

$$\underline{R}(Y)=Y\{X\in U/R/X\subseteq Y\} \tag{1}$$

$$\overline{R}(Y)=Y\{X\in U/X\cap Y\neq\phi\} \tag{2}$$

定义3 $C,D\subseteq A$分别为条件属性集和决策属性集,$ind(C),ind(D)$表示由C,D决定的不可区分关系,关系$ind(C)$的等价类的集合称为条件类,用U/C表示,关系$ind(C)$的等价类的集合称为决策类,用U/C 表示.则定义D的C正域是U中所有根据分类U/C的信息准确地划分到关系U/C的等价类中的对象集合,即

$$pos_C(D)=Y_{Y\in U/D}\ \underline{ind(C)}Y \tag{3}$$

定义4 设$\phi\subset X\subseteq C,\phi\subset Y\subseteq D,U/Y\neq\{U\}$ 定义Y的关于X的支持子集为

$$S_X(Y)=Y_{W\in U/Y}(Y_{V\in U/X,V\subseteq W}V) \tag{4}$$

定义5 设$\phi\subset X\subseteq C,\phi\subset Y\subseteq D,U/Y\neq\{U\}$.定义$Y$的关于$X$的支持度为

$$SPT_X(Y)=|S_X(Y)|/|U| \tag{5}$$

定义6 设$\phi\subset X\subseteq C,\phi\subset Y\subseteq D,U/Y\neq\{U\}$.给定$x\in X$,定义$x$在$X$中的重要度(相对于$Y$而言)为

$$\begin{aligned}sig^Y_{X-|x|}(x)&=(|S_X(Y)|-|S_{X-|x|}(Y)|)/|U|\\&=SPT_X(Y)-SPT_{X-|x|}(Y)\end{aligned} \tag{6}$$

特别的,当$X=\{x\}$,则

$$\begin{aligned}sig^Y_{X-|x|}(x)&=sig^Y_\phi(x)\\&=(|S_X(Y)|-|S_{X-|x|}(Y)|)/|U|\\&=|S_x(Y)|/|U|=SPT_X(Y)\end{aligned} \tag{7}$$

如果$sig^Y_{X-|x|,\alpha}(x)>0$,则称$x$在$X$中是重要的;如果$sig^Y_{X-|x|,\alpha}(x)=0$,则称$x$在$X$中是不重要的.

1.2 基于样本信息熵的相关概念

定义7 临时表DT^*.根据给定的决策表DT构造的一个新的决策表,其中每一行代表了决策表中决策值不同的待区分的样本对,每一列代表一个候选划分点,若依据划分点划分可以辨别某个样本对,则划分点对应的区间变量值为1,否则,变量值为0.

定义8 断点p的重要度.以文献《基于粗集理论的数据离散化方法》中的选择概率来定义断点的重要度.M行N列临时表DT.任一候选断点的重要度由其行向选择概率和列向选择概率按下面的公式定义:

$$ID_p=(\min(M,N))\times SP_r+\max(M,N)\times SP_c)/(M+N) \tag{8}$$

其中,SP_r表示行向选择概率,SPc表示列向选择概率.

定义9 属性a的重要度.设结果断点子集中包含属性a的断点数目为m,则属性a的重要度为

$$ID_a=\sum_{i=1}^{m}ID_{pi} \tag{9}$$

定义10 样本决策属性值j的概率P_j.设条件属性的个数为n_a,满足决策属性值为j的条件属性的个数为n_j,则:

$$P_j=\sum_{k=1}^{n_j}ID_{ak}\Big/\sum_{k=1}^{n_a}ID_{ak} \tag{10}$$

定义11 样本x的信息熵$H(x)$.设样本决策属性值j的概率为P_j,则:

$$H(x) = -\sum_{j=1}^{d} p_j \log_2 p_j \tag{11}$$

其中,d为决策种类个数,j为决策属性值.

定义12　本文用到的数据类型说明.

正常数据:理论上是正确的数据;

噪声数据:一部分条件属性的值偏离正常属性值范围;

错误数据:所有条件属性值没有偏离正常属性值范围,但决策属性值错误;

异常数据:所有的条件属性值均偏离正常属性值范围.

2　决策基于属性重要性和样本信息熵的数据离散化后处理算法

算法思想:首先对决策表进行离散化(结果可能删除一部份属性);然后根据粗糙集理论计算决策表中属性的重要度,删除重要度为0的属性;最后计算每个样本的信息熵,根据信息熵的大小判断样本的数据类型.

按照定义5计算出样本　所有决策属性值的概率,其中最大概率对应的决策值为α_0,样本的实际决策属性值为α_1.给定样本x的信息熵阈值$\beta(0<\beta<1)$,则基于属性重要性和样本信息熵的数据离散化后处理算法如下:

(1)采用文献[1]对决策表进行离散化.

(2)按照公式(7)计算决策表各个条件属性的重要度,删除重要度为0的属性.

(3)按照公式(Ⅱ)计算每个样本的信息熵.

(4)判断决策表中每个样本x的数据类型.

①正常数据　如果$H(x)=0$,则样本x为正常数据.

②噪声数据　如果$H(x)=\beta$,且$\alpha_0=\alpha_1$,则样本x为噪声数据.

③错误数据　如果$H(x)>\beta$,且$\alpha_0\neq\alpha_1$,则样本x为错误数据.

④异常数据　如果$H(x)\leq\beta$,则样本x为异常数据.

(5)结束

3　仿真实验讨论

3.1　评价方法

(1)正确率　设样本总数为n,正确识别样本个数为n_1,则识别正确率为$p_1=n_1/n$.

(2)错误率　设样本总数为n,错误识别样本个数为n_2,则识别错误率为$p_2=n_2/n$.

3.2　实验

3.2.1　实验数据库

实验中使用的数据库是UCI中常见的数据库,见表1.对UCI数据库中的iris, austra, heart, pima四个数据库人为重新构造,构造思想如下:为每个库的每个类别按照定义12构造新的记录数据.每个类的构造方法如下:

把原来每个类的数据分成4等份,其中第一份保留作为正常数据;第二份的大多数属性的值改为偏离原来属性正常值的范围,把它们作为噪声数据;第三份是不改变原来的条件属性值,但把决策属性值改为其它类别,把它们作为错误数据;第四份改变所有的条件属性值,但不改变决策属性值,把它们作为异常数据.实验中,$\beta=0.45$.

表1　实验中用到的数据库

Table 1　The date base in experiment

名称	样本数	条件属性数	决策类数
Iris	150	4	3
austra	690	14	2
heart	270	9	2
pima	768	8	2

3.2.2 实验结果

表2 四种类型数据库的正确率和错误率

Table 2 Fouv type datebases corvect and error rate

	正常数据		噪声数据		错误数据		异常数据	
	正确率(%)	错误率(%)	正确率(%)	错误率(%)	正确率(%)	错误率(%)	正确率(%)	错误率(%)
iris	96.0	4.0	92.1	7.9	96.5	3.5	96.6	3.4
austra	87.7	12.3	81.2	18.8	89.3	10.7	90.0	10.0
heart	77.4	22.6	72.4	27.6	83.5	16.5	85.1	14.9
pima	70.6	29.4	68.2	31.8	78.6	21.4	81.0	19.0

观察表2可以看出,对于正常数据,iris的准确率最高,错误率最低、austra、heart和pima的准确率依次降低,错误率依次上升,各个数据集的准确率依次相差8.3%、10.3%和6.8%.噪声数据、错误数据和异常数据的正确率和错误率也基本上是按照上述顺序稳定排序.分析原因可能是与各个数据库的本身的数据特点有关,iris数据库不同属性的均值差别较小,每种类型记录个数相等,不同类型数据分布比较均匀,而austral、heart和pima数据库的不同属性的均值差别依次变大,属性值的偏差也依次变大,决策种类的记录数分布依次变得更不均匀.结果表明本文的算法与数据本身的特点相关,对于不同特点的数据判断效果有很大差别.

4 结论

本文讨论了基于粗糙集理论属性重要性的相关概念,提出了样本信息熵的定义,并在此基础上提出了基于属性重要性和样本信息熵的数据离散化后处理算法.实验结果表明本文算法是有效的,但对于不同特点的数据判断效果有很大差别.

今后的研究工作包括:(1)改进此算法,使改进的算法能够处理不同特点的数据,尽可能提高算法的普遍适用性.(2)结合遗传算法,进一步提高解决大规模数据问题的效率.

参 考 文 献

〔1〕Nguyen H S, Skowron A. Quantization of real values attributes, rough set and Boolean reasoning approaches: proc. of the 2nd Joint Annual Conference on Information Science〔C〕. Wrightsville Beach:NC,1995.34-37.

〔2〕Nguyen S H, Nguyen H S. Some efficient algorithms for rough set methods: Proc. of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems〔C〕. Spain: Granada, 1996.1451-1456.

〔3〕Susmaga R. Analyzing discretizations of continuous attributes given a monotonic discrimination function〔J〕. Intelligent Data Analysis, 1997,1(4):157-179.

〔4〕Dai Jian-Hua, Li Yuan-Xiang. Study on discretization based on rough set theory: Proc. of the first International Conference on Machine Learning and Cybernetics〔C〕.Beijing:[s.n.], 2002. 1371-1373.

〔5〕王国胤. Rough集理论与知识获取〔M〕. 西安: 西安交通大学出版社, 2001.

〔6〕Chen Cai-Yun, Li Zhi-Guo, Qiao Sheng-Yong, et al. Study on discretization in rough set based on genetic algorithm: Proc. of the Second International Conference on Machine Learning and Cybernetics〔C〕.Xi′ an: [s.n.], 2003: 1430-1434.

〔7〕Huang Jin-Jie, Li Shi-Yong. A GA-based approach to rough data model: Proc. of the 5th World Congress on Intelligent Control and Automation〔C〕.Hangzhou:[s.n.],2004.1880-1884.

〔8〕侯利娟,王国胤,聂能,等. 粗糙集理论中的离散化问题〔J〕. 计算机科学, 2000, 27(12): 89-94.

〔9〕何亚群, 胡寿松. 粗糙集中连续属性离散化的一种新方法〔J〕. 南京航空航天大学学报, 2003, 35(3): 213-215.

〔10〕李兴生. 一种基于云模型的决策表连续属性离散化方法〔J〕. 模式识别与人工智能, 2003, 16(3): 33-38.

〔责任编辑 郑 瑛〕

阅读此文的还阅读了：

1. 基于信息熵的决策属性分类挖掘算法及应用
2. 用WebBrowser控件创建Web浏览器
3. 连续属性离散化的MaxDiff方法
4. MAPI高级邮件编程机制
5. VFP中的表格控件
6. Web浏览器在VB下的实现
7. 用VB5.0制作工具条
8. WEB应用开发中的代码重用技术
9. 用面向对象的程序设计(OOP)方法设计通用子类建造MIS
10. 职业技术院校“面向对象”程序设计课程教学思考
11. 判断文件夹的内容大小
12. OLE技术在应用型地理信息系统开发中的应用
13. 基于竞争型网络的连续属性离散化方法
14. 一种基于条件熵的粗糙集连续属性离散化方法
15. 利用Delphi开发E-mail发送程序
16. 基于杂度削减的连续属性离散化方法

17. 基于改进遗传算法的连续属性离散化方法

18. 选择题控件在VB中的设计与实现

19. JavaScript的面向对象特性浅析与范例

20. 用CDONTS组件开发邮件发送程序

21. 用列表框制作动态畅销书榜单

22. 语言学的科学属性及其研究方法的来源与选择

23. 面向对象技术在过程仿真中的应用

24. 一种基于SOFM网络的连续属性离散化方法

25. 基于遗传算法的连续属性离散化方法研究

26. 中小型企业账务管理系统的设计与实现

27. 基于数据分区的连续属性整体离散化方法研究

28. Visual Basic封装技术的实现

29. 局域网信息发布程序的设计与实现

30. JavaScript函数与事件应用

31. ActiveX技术在海洋资料自动采集系统研制中的应用

32. 基于对象分布的连续属性离散化方法

33. 用VB开发采样趋势曲线显示软件

34. 一种基于条件熵的粗糙集连续属性离散化方法

35. 试论菜肴的“香”

36. 基于连续属性离散化和SVM的分类预测方法

37. 基于Visual Basic 2008的Access数据库类的设计

38. 基于信息熵理论的连续属性离散化方法

39. 一种基于进化算法的连续属性离散化方法

40. VB6.0中橡皮筋特征绘图的实现方法

41. 物流信息化过程中基于熵的信息组织分析方法研究

42. ActiveX控件在Exce197中的应用

43. 粗糙集理论中基于属性重要性的离散化方法

44. 浅论公证程序证据审查

45. 一种基于遗传算法的连续属性离散化方法

46. 用Delphi实现工控机与单片机系统的串行通信

47. C~(++)Builder中用户自定义组件的实现

48. 快速入门ActionScript脚本语言教学方法探究

49. 基于信息熵的粗糙集连续属性离散化算法

50. 一种新的基于连续属性离散化的属性约简方法