面向属性值遗漏数据决策树分类算法研究

邱云飞[1] 李 雪[1] 王建坤[1] 邵良杉[2]
(辽宁工程技术大学软件学院 葫芦岛 125100)[1] (辽宁工程技术大学系统工程研究所 葫芦岛 125100)[2]

摘 要 在已有的多种决策树测试属性选择方法中,未见将属性值遗漏数据处理集成在测试属性选择过程中的报道,而现有的属性值遗漏数据处理方法都会不同程度地带入偏置。基于此,提出了一种将基于联合熵的信息增益率作为决策树测试属性选择标准的方法,用以在生成决策树的过程中消除值遗漏数据对测试属性选择的影响。在WEKA机器平台上进行了对比实验,结果表明,改进算法能够从总体上提高算法的执行效率和分类精度。

关键词 属性值遗漏数据,联合熵,决策树
中图法分类号 TP311 **文献标识码** A

Research on the Missing Attribute Value Data-oriented Decision Tree

QIU Yun-fei[1] LI Xue[1] WANG Jian-kun[1] SHAO Liang-shan[2]
(School of Software, Liaoning Technical University, Huludao 125100, China)[1]
(System Engineering Institute, Liaoning Technical University, Huludao 125100, China)[2]

Abstract In the existing multiple choice methods of decision tree's test attributes, can't see such report as "Let missing data processing integrated in the selection process of test attributes", however, the existing process methods of missing attribute value data could draw into bias in different degrees, based on this, proposed an information gain rate based on combination entropy as the decision tree's testing attributes selection criteria, which can eliminate missing value arrtibutes' infulence on testing attributes selection, and carry out contrast experiments on WEKA. Experiment results indicate that the improvement can significantly increase whole efficiency and classification accuracy of the algorithm operation.

Keywords Missing attribute value data, Combination entropy, Decision tree

属性值遗漏数据在现实世界中频繁出现,数据集中一个对象遗漏一个或多个属性值并不少见[1]。一个对象虽然遗漏了一个或多个属性值,但在其他属性值上仍然对决策树分类器的构造提供着有价值的信息。因此,研究有效利用值遗漏数据训练决策树的方法具有一定的实际意义。

现有的决策树生成算法在选择测试属性时都是利用基于属性的现有值去估计遗漏值的策略来确定测试属性的选择标准的[1,2],比如John Mingers[3]按现有值在整个数据集中出现的比例来估计遗漏值。Safavian & Landgrebe[4]用现有值中的最常见值来估计遗漏值,这些方法虽然可以在一定程度上解决属性值遗漏数据对测试属性的选择,但都具有一定的主观成分,更符合客观认识规律的测试属性选择标准仍是决策树领域的研究目标之一。

决策树分类算法中最核心的功能是如何选择测试属性[5]。C4.5[2,6]采用了Shannon熵为基础的信息增益率来选择测试属性,而Shannon熵的计算方法是以数据集中没有属性值遗漏数据为前提的,所以当数据集中含有属性值遗漏数据时将引入偏置。本文介绍了一种能够处理属性值遗漏数据的基于粗糙集理论的联合熵,并利用其构造测试属性选择标准,以达到消除偏置的目的。

1 算法提出

在生成决策树的过程中,为了消除值遗漏数据对测试属性选择的影响,提出了新的算法MultiInfoTree。

1.1 算法中测试属性的选择标准

基于联合熵的互信息[7]能够很好地处理属性值遗漏数据,而且不引入任何偏置信息,因此可利用其构造测试属性的选择标准。

$S=(U,AT)$是信息系统。其中U是对象的非空有限集合,AT是属性的非空有限集合。对于$\forall a\in AT$有$a:U\to V_a$,其中V_a称为a的值域。如果$\exists a\in AT$使得V_a含有空值,则称S为一个不完备信息系统,否则它是完备的,用$*$来表示空值。

在不完备信息系统[2] $S=(U,AT)$中,$U/SIM(A)=\{S_A(u_1),S_A(u_2),\cdots,S_A(u_{|U|})\}$,$A$的联合熵被定义[5]为

$$CE(A)=\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{C_{|U|}^2-C_{|S_A(u_i)|}^2}{C_{|U|}^2},i\leqslant|U| \tag{1}$$

当$S_1=(U,P)$,$S_2=(U,Q)$是两个不完备信息系统时,P

到稿日期:2010-11-19 返修日期:2011-03-27 本文受国家自然科学基金(70971059),辽宁省创新团队项目(2009T045)和辽宁省科技攻关项目(2007308003)资助。

邱云飞(1976—),男,博士,副教授,主要研究方向为数据挖掘理论与应用,E-mail:lntulixue17@163.com;李 雪(1987—),女,硕士生,主要研究方向为数据挖掘;王建坤(1985—),男,硕士生,主要研究方向为数据挖掘和文本分类;邵良杉(1961—),男,博士生导师,主要研究方向为数据挖掘。

和Q的互信息[7]定义为

$$CE(P;Q)=CE(P)+CE(Q)-CE(P\cup Q) \quad (2)$$

$P\cup Q$的联合熵定义为

$$CE(P\cup Q)=\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{C_{|U|}^2-C_{|S_P(u_i)\cap S_Q(u_i)|}^2}{C_{|U|}^2} \quad (3)$$

根据式(2),对于不完备决策表$S=(U,C,d,V,f)$,即生成决策树的训练集,有$c\in C$和d的互信息为

$$CE(\{c\};\{d\})=CE(\{c\})+CE(\{d\})-CE(\{c\}\cup\{d\}) \quad (4)$$

式中,数据对象在c上的取值可能遗漏,而d是类别属性,不存在遗漏值。将取$CE(\{c\};\{d\})$最大的属性c为当前结点的测试属性。

为了简化选择测试属性的计算,对式(4)进行了分析。根据式(2)和式(3),有

$$\begin{aligned}&CE(\{c\};\{d\})\\&=\frac{1}{|U|}\sum_{i=1}^{|U|}\{\frac{C_{|U|}^2-C_{|S_c(u_i)|}^2}{C_{|U|}^2}+\frac{C_{|U|}^2-C_{|S_d(u_i)|}^2}{C_{|U|}^2}-\\&\frac{C_{|U|}^2-C_{|S_c(u_i)\cap S_d(u_i)|}^2}{C_{|U|}^2}\}\\&=\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{C_{|U|}^2-C_{|S_c(u_i)|}^2-C_{|S_d(u_i)|}^2+C_{|S_c(u_i)\cap S_d(u_i)|}^2)}{C_{|U|}^2}\end{aligned} \quad (5)$$

对于任意u_i来说,u_i在属性c上的取值有两种情况:$*$值或非$*$值,而u_i在属性d上的取值不能遗漏。因此,分以下几种情况进行讨论:

当$c(u_i)=*$时,$S_c(u_i)=U$,

所以$\frac{C_{|U|}^2-C_{|S_c(u_i)|}^2-C_{|S_d(u_i)|}^2+C_{|S_c(u_i)\cap S_d(u_i)|}^2)}{C_{|U|}^2}=0$。就是说,当计算$CE(\{c\};\{d\})$时可以不考虑$c(u_i)=*$的数据对象。

当$c(u_i)\neq *$时,即可直接求$CE(\{c\};\{d\})$,只是$CE(\{c\};\{d\})$中的$S_d(u_i)$退化为u_i在决策属性d上的等价类$IND_d(u_i)$。$S_c(u_i)$的计算也非常简便,即为$c(u_i)$值相等的数据对象和$c(u_i)=*$的数据对象的并集。如果属性c值域中有k个值,则U被分为k个这样的子集合,将其称为极大相容块[8],用$MCB_c(u_i)$表示。

设不完备决策表$S=(U,C,d,V,f)$中在属性c上值遗漏的数据对象有l个,那么在计算$CE(\{c\};\{d\})$时只需要计算$|U|-l$个$c(u_i)\neq *$的数据对象,因此有

$$\begin{aligned}&CE(\{c\};\{d\})=\frac{1}{|U|}\sum_{i=1}^{|U|-l}\{\\&\frac{C_{|U|}^2-C_{|S_c(u_i)|}^2-C_{|S_d(u_i)|}^2+C_{|S_c(u_i)\cap S_d(u_i)|}^2}{C_{|U|}^2}\}=\frac{1}{|U|}\sum_{i=1}^{|U|-l}\{\\&\frac{C_{|U|}^2-C_{|MCB_c(u_i)|}^2-C_{|IND_d(u_i)|}^2+C_{|MCB_c(u_i)\cap IND_d(u_i)|}^2}{C_{|U|}^2}\}\end{aligned} \quad (6)$$

基于联合熵的互信息作为测试属性的选择标准与基于shannon熵的互信息作为测试属性的选择标准都有偏向属性值较多属性的问题。采用与C4.5相同的方法也将信息增益率[2] $gain_ratio(c)$作为最终的属性选择标准,即

$$gain\_ratio(c)=\frac{CE(\{c\};\{d\})}{CE(c)} \quad (7)$$

1.2 算法描述

面向不确定性数据的决策树分类算法*MultiInfoTree*:

输入:训练数据集*samples*;候选属性的集合*attribute_list*输出:一棵决策树

(1)创建节点N;

(2)if *samples* 中数据对象都在同一类C中 then

(3)返回N作为叶节点,以类C标记;

(4)if *attribute_list* 为空 then

(5)返回N作为叶节点标记为*samples*中最普通的类;

(6)选择*attribute_list*中具有最高信息增益比率的属性*test_attribute*;

(7)标记节点N为*test_attribute*;

(8)for each *test_attribute* 中的已知值e_i//划分 *samples*

(9)由节点N长出一个条件为*test_attribute*=e_i的分枝;对*test_attribute*属性上每个值遗漏数据对象在*samples*中求相似集,由相似集中的大多数数据对象的类别来决定值遗漏数据对象的分支;

(10)设S_i是*samples*中*test_attribute*=e_i的样本的集合;//一个划分

(11)if S_i为空 then

(12) 加上一个树叶,标记为*samples*中最普通的类;

(13) else 加上一个由*MultiInfoTree*(S_i,*attribute_list*,*test_attribute*),*test_attribute*)返回的节点。

2 与其他决策树构造算法的比较

为验证本文算法的性能,进行了模拟实验,并分析了实验的结果。整个实验情况如下所述。

为了与其他算法做比较,这里决定采用WEKA[9]自带的J48算法(即C4.5算法)和本文的MultiInfoTree算法。所用实验数据为UCI数据库(http://www.iCs.uCi.edu/~mlearn/)中的7个数据集。

2.1 实验结果

实验环境:IntelPentium 4处理器,1GB内存。操作系统:WindowsXP。算法利用JAVA编程实现。

在WEKA机器学习平台上,利用算法J48和MultiInfoTree分别对数据集进行分类,比较树的规模、建树时间等多个方面,所得结果如表1所列。

表1 数据库利用算法J48和MultiInfoTree分类比较表

数据库名称	样本总数	条件属性个数	样本缺失率	叶子节点数		树的规模		建树时间		分类精度	
				J48	MIT	J48	MIT	J48	MIT	J48	MIT
bridges_version2.arff	107	13	69.23%	10	21	14	29	0.45s	0.02s	56.1905%	71.4286%
anneal.ORIG.arff	898	39	74.36%	56	38	68	49	4.47s	0.7s	87.9733%	92.3163%
soybean.arff	683	36	94.44%	61	82	93	126	2.44s	0.25s	87.5081%	88.5798%
agaricus-lepiota.arff	8123	23	4.35%	32	48	47	65	4.2s	3.33s	63.0432%	62.3169%
audiology.arff	226	70	10.00%	32	18	54	27	2.61s	0.11s	78.5664%	77.8761%
breast-cancer.arff	286	10	20.00%	4	6	6	10	1.33s	0.02s	75.5245%	74.1259%

注:表1中样本缺失率的计算方式为:含有缺失值的属性个数/总属性个数;"MIT"为算法MultiInfoTree。

为方便比较,对不含缺失值的数据集car.arff(Number of Instances:1728, Number of Attributes:6, Missing Attribute

Values:none)进行处理，当其缺失率分别为10%，20%，30%，40%，50%，60%，70%，80%，90%时，利用两算法分别对数据集进行分类，得出如图1所示的实验结果。

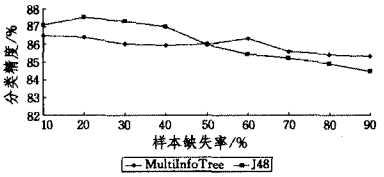

图1 算法J48、MultiInfoTree在不同样本缺失率下分类精度比较

2.2 结果分析

比较表1中两种算法分类结果可以看出，分类所得结果有明显不同，原因在于MultiInfoTree算法和J48算法对属性进行分割的标准不同，它们分别选用联合熵和Shannon熵作为分割标准。改进后的算法MultiInfoTree效率提高10～20倍。其原因是联合熵的计算比Shannon熵要简单得多，避免了计算对数log消耗的时间，改进算法的执行效率明显优于C4.5。

从图1的实验结果可以看出，随着数据样本缺失值的增加，J48算法的分类精度下降明显，而MultiInfotree算法的分类精度未有很大变化。这说明J48算法在处理属性值遗漏数据时会带入偏置，而MultiInfotree算法在生成决策树的过程中能够消除值遗漏数据对测试属性选择的影响。当缺失率大于50%时，改进后的算法的分类精度提高最为明显，这说明MultiInfoTree算法较适合于缺失率在这个区间的数据集。

结束语 在实际的数据中，属性值遗漏数据是无处不在的。本文把基于联合熵的信息增益率作为决策树测试属性选择的标准，它能够在生成决策树的过程中消除值遗漏数据对测试属性选择的影响，更适合于实际数据。最后，通过实验数据验证了MultiInfoTree算法能够从总体上提高算法执行效率和分类精度，非常适合于样本缺失率大于50%的数据集的分类问题。当样本缺失率在其他区间时，将通过组合分类器[10]的技术来提高分类准确率。我们会在以后的工作中不断扩展和细化这个领域的研究，使之更加完善。本文研究成果解决了从大规模、不确定性数据集中发现决策树分类模型的问题，为应用决策树分类技术提供了更为广阔的空间。

参 考 文 献

[1] Gustavo E A, Batista P A, Monard M C. An Analysis of Four Missing Data Treatment Methods for Supervised Learning [J]. Applied Artificial Intelligence, 2003, 17(5/6): 519-533

[2] Kryszkiewicz M. Rules in incomplete information systems [J]. Information Sciences, 1999, 113: 271-292

[3] Mingers J. An empirical comparison of selection measures for decision-tree induction [J]. Machine Learning, 1989, 3(4): 319-342

[4] Safavian S R, Landgrebe D. A Survey of Decision Tree Classifier Methodology[R]. 47907. School of Electrical Engineering, Purdue University, 1991: 1-58

[5] 冯少荣. 决策树算法的研究与改进[J]. 厦门大学学报：自然科学版, 2007, 20 (4): 498-500

[6] Quinlan J R. C4.5: Programs for Machine Learning [S]. Morgan Kaufman, 1993

[7] Qian Yunhua, Liang Jiye. A new method for measuring the uncertainty in incomplete information systems [J]. International Journal of Uncertainty. Fuzziness and Knowledge-Based Systems, 2008(9)

[8] Leung Y, Li D Y. Maximal consistent block technique for rule acquisition in incomplete information systems [J]. Information Science, 2003, 153: 85-106

[9] 赵蕊. 基于WEKA平台的决策树算法设计与实现[D]. 长沙：中南大学, 2007: 43-46

[10] 旷海兰，罗可，刘新华，等. 一种基于粗糙集理论的组合分类器构造方法[J]. 计算机工程与应用, 2006, 16

(上接第168页)

阶段的工作流程以及结构关系。然后，给出基于Hadoop的并行k-means算法设计时需要思考的主要问题、算法设计的主要流程以及方法和策略等。最后，通过在多组不同大小数据集上的实验表明，我们设计的并行聚类算法PKMeans适合运行于大规模云计算平台，可以有效地应用于实际中海量数据的分析和挖掘。

随着云计算概念的兴起，基于云计算平台的数据挖掘、聚类算法的研究逐渐成为国内外学者的研究热点。未来的研究方向包括：1) 研究聚类算法并行化的一般规律，找到数据规模、算法复杂性、节点数之间的关系，发现加速比和可扩展性的影响因素，从而设计出高效的并行聚类算法；2) 研究基于云计算平台的数据挖掘应用中的信息安全和隐私保护等问题，该问题的解决对于云计算在实际商务中的应用将起到关键性的作用。

参 考 文 献

[1] Han J W, Kamber M. Data mining: concepts and techniques [M]. San Francisco, US: Morgan Kaufmann, 2001

[2] Buyya R, Yeo C S, Venugopal S. Market-oriented cloud computing: vision, hype, and reality for delivering IT services as computing utilities, Keynote Paper [C] // Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications. Dalian, China, 2009: 25-27

[3] Armbrust M, Fox A. Above the clouds: a Berkeley view of cloud computing[R]. USA: University of California at Berkeley, 2009

[4] Erdogmus H. Cloud computing: does nirvana hide behind the nebula[J]. IEEE Software, 2009, 26(2): 4-6

[5] 郑纬民. 云计算的大幕已经拉开[J]. 中国计算机学会通讯, 2009, 2(6): 6-7

[6] Ghemawat S, Gobioff H, Leung S. The google file system[J]. S ACM SIGOPS Operating Systems Review, 2003, 37(5): 29-43

[7] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters [C] // Proceedings of Operating Systems Design and Implementation. San Francisco, CA, 2004: 137-150

[8] Xu X W, Jager J, Kriegel H P. A fast parallel clustering algorithm for large spatial databases [J]. Data Mining and Knowledge Discovery, 1999, 3(3): 263-290