基于粗糙集理论的不完备数据填补方法

孟 军[1,2],刘永超[2],莫海波[1]
MENG Jun[1,2],LIU Yong-chao[2],MO Hao-bo[1]

1.大连理工大学 计算机科学与工程系,辽宁 大连 116023
2.大连理工大学 软件学院,辽宁 大连 116620
1.Department of Computer Science and Engineering,Dalian University of Technology,Dalian,Liaoning 116023,China
2.School of Software,Dalian University of Technology,Dalian,Liaoning 116620,China
E-mail:mengjun@dlut.edu.cn

MENG Jun,LIU Yong-chao,MO Hao-bo.New method of packing missing data based on rough set theory.Computer Engineering and Applications,2008,44(6):175-177.

Abstract: ROUSTIDA has highly ability of packing missing data.But it still has some incomplete information.This paper takes advantage of the discrimination of attributes suggested from extended discriminable matrix.The improved ROUSTIDA extends the ability of packing missing data,and has a new ability to eliminate noise data.It also reduces running time.All that has been proved in experiments.

Key words: rough set;similarity relation;extended discriminable matrix

摘 要:ROUSTIDA 算法具有较好的数据填补能力,但依然会出现一些不完备信息。利用了可扩充辨识所反映的对象间的属性差异信息,对遗失属性进行填充,从而使改进后的 ROUSTIDA 算法的填充能力得到了很大的改善,同时还具备了初步排除噪声数据的能力,在性能上也有了很大的提高,实验表明改进的算法具有很好的实用价值。

关键词:粗糙集;相似关系;扩充辨识矩阵
文章编号:1002-8331(2008)06-0175-03 **文献标识码:**A **中图分类号:**TP311

1 引言

在信息化和数字化日益普及的今天,各行各业中都积累了大量的历史数据,人们采用各种各样的数据挖掘技术从这些历史数据中找出自己感兴趣的信息,来指导人们以后的决策。但是有很多数据是不完善的,这些遗失的数据影响了随后的数据分析,必须采取一种可行的数据填补方法,在保证使填补后的数据产生的分类规则具有尽可能高的支持度且在规则集中的前提下,对数据进行填补。目前填补遗失值的方法有均值法、最大频法[1]、不完备数据分析方法(ROUSTIDA)[2]等,近来又有不少学者提出了新的填补方法,如基于信息表断点的填充方法[3]。由于粗糙集理论中的 ROUSTIDA 算法是现有填补算法中填补性能最好的[4,5],本文在遵循原算法指导思想的前提下对使用范围及填充能力进行了改进,实验证明改进后的算法在填充遗失属性时更合理更有效。

2 ROUSTIDA 方法

2.1 ROUSTIDA 算法描述

不完备信息系统中的遗失数据值的填补,应该尽可能反映此信息系统所体现的基本特征以及隐含的内在规律。ROUSTIDA 算法的基本思想是:遗失数据值的填充应是完整化后的信息系统产生的分类规则具有尽可能高的支持度,产生的规则尽可能集中。下面是算法的描述:

输入:不完备信息系统 $S^0=\langle U^0,A,V,f^0\rangle$;

输出:完备的信息系统 $S^r=\langle U^r,A,V,f^r\rangle$。

步骤 1 计算初始可辨识矩阵 $\boldsymbol{M}^0$,$\boldsymbol{MAS}_i^0$ 和 $\boldsymbol{MOS}^0$,令 $r=0$;

步骤 2

首先,对于所有的 $i\in MOS^r$,计算 NS_i^r;

其次,产生 S^{r+1};

(1)对于 $i\notin MOS^r$,有 $a_k(x_i^{r+1})=a_k(x_i^r)$,$k=1,2,\cdots,m$;

(2)对于所有的 $i\in MOS^r$,对所有 $k\in MAS_i^r$ 作循环;

①如果 $|NS_i^r|=1$,设 $j\in NS_i^r$,若 $a_k(x_j^r)=*$,则 $a_k(x_i^{r+1})=*$;否则 $a_k(x_i^{r+1})=a_k(x_j^r)$;

②否则:

(i)如果存在 j_0 和 $j_1\in NS_i^r$,满足 $(a_k(x_{j0}^r)\neq *)\cap(a_k(x_{j1}^r)\neq *)\cap(a_k(x_{j1}^r)\neq a_k(x_{j0}^r))$,则 $a_k(x_i^{r+1})=*$;

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60373095)。

作者简介:孟军(1964-),女,副教授,主要研究领域为数据库与数据挖掘;刘永超(1981-),男,硕士研究生,主要研究领域为文本挖掘技术;莫海波(1983-),男,硕士研究生,主要研究领域为文本挖掘技术。

收稿日期:2007-06-14 **修回日期:**2007-08-13

(ii)否则,如果存在 $j_0 \in NS_i^r$,满足 $a_k(x_{j0}^r) \neq *$,则 $a_k(x_i^{r+1}) = a_k(x_{j0}^r)$;

(iii)否则 $a_k(x_i^{r+1}) = *$;

最后,如果 $S^{r+1} = S^r$ 结束循环转步骤 3;否则,重新计算 M^{r+1},MAS_i^{r+1} 和 MOS^{r+1};$r = r+1$;转步骤 2;

步骤 3 如果信息系统还有遗失值,可用取属性值中平均值(数字型)或最多出现值(符号型)的方法处理(当然,也可以用其他方法);

步骤 4 结束。

2.2 ROUSTIDA 算法分析

首先,从填补能力上分析,ROUSTIDA 算法本身是用相似关系代替不可分辨关系,经过算法填充后的信息系统产生的分类规则具有尽可能高的支持度,同时产生的规则也尽可能的集中。但是在算法中也存在着一些不太令人满意的地方,在步骤 2->2->(2)->②->(i)中,假设数据 i 的无差别对象集中存在多于两条数据,遗失属性为 a_k,j_0,j_1 是其中的两条,且 j_0,j_1 在属性 a_k 上的属性值都不空,并且不相等,按照算法处理方法这条数据就不被填充,如表 1 所示。

表 1 ROUSTIDA 不能填充属性值的情况

U	a_1	a_2	a_k	a_4
i	yes	no	*	no
j_0	yes	no	yes	*
j_1	yes	no	no	no

算法结束后生成的信息系统还是不完备的,还得借助于其他算法来把这些空缺值给补上。

其次,从时间复杂度上分析,ROUSTIDA 的时间复杂度是 $O(n*k*p)$,其中 p 是可能的迭代的次数。对于大数据量来说,每次填充完后再借助于别的填补方法,增加了整个信息处理系统的负担;同时,在每轮循环结束后都要重新计算扩充辨识矩阵 $\boldsymbol{M}$,这都对时间性能的改善造成很大的影响。

再次,从区分不完备噪声数据能力上分析,该算法无法区分出不完备的噪声数据。

最后,由于原来的 ROUSTIDA 算法无法填充一些属性值,还需要借助步骤 3 中的其他的填充算法,如均值,这样有可能使填充后的系统出现噪声数据。

3 改进的 ROUSTIDA 算法

因为原算法是对扩充辨识矩阵进行操作,当数据中有很多相同的数据时,就会影响算法的时间性能,所以改进后的算法,在对遗失数据进行填补前,先对扩充辨识矩阵进行化简。即相同的数据在矩阵中只出现一次,同时保存它在信息系统中出现的次数,这样可以很大程度上减少扩充矩阵的维数,减轻了后面算法的运算量。

改进后的 ROUSTIDA 在原算法的基础上,就原算法步骤 2 中,对一个遗失属性有两个不同的相似对象且这两个对象的这个属性值不同的情况下不能填充数据的缺陷进行了改进。因为要保证填补后的信息系统产生的规则有尽可能高的支持度,所以改进的算法从相似对象中找规则支持度最高属性值来填充,这样既保证了遗失属性的填补,又保证了填补的数据满足原算法的基本思想,同时还去除了原算法可能产生噪声数据的潜在问题。

3.1 改进后的算法描述

输入:不完备信息系统 $S^0 = \langle U^0, A, V, f^0 \rangle$;

输出:完备的信息系统 $S^r = \langle U^r, A, V, f^r \rangle$。

步骤 1 计算初始可辨识矩阵 $\boldsymbol{M}^0$,$\boldsymbol{MAS}_i^0$ 和 $\boldsymbol{MOS}^0$,令 $r=0$;

步骤 2

首先,对于所有的 $i \in MOS^r$,计算 NS_i^r;

其次,产生 S^{r+1};

(1)对于 $i \notin MOS^r$,有 $a_k(x_i^{r+1}) = a_k(x_i^r)$,$k=1,2,\cdots,m$;

(2)对于所有的 $i \in MOS^r$,对所有 $k \in MAS_i^r$ 作循环;

①如果 $|NS_i^r|=1$,设 $j \in NS_i^r$,若 $a_k(x_j^r) = *$,则 $a_k(x_i^{r+1}) = *$;否则 $a_k(x_i^{r+1}) = a_k(x_j^r)$;修改 M^r 的 $M(i,j)$和 $M(j,i)$同时修改 MAS_i,MOS,NS_i 转步骤 2;

②否则:

(i)对 $j \in NS_j^r$ 做循环,同时在 M^r 中统计属性 k 在第 j 行中出现的次数,记录出现次数最少,且这个属性值不为遗失值的无差别对象 J,如果存在 j_0 和 $j_1 \in NS_i^r$,满足 $(a_k(x_{j0}^r) \neq *) \cap (a_k(x_{j1}^r) \neq *) \cap (a_k(x_{j1}^r) \neq a_k(x_{j0}^r))$ 则 $a_k(x_i^{r+1}) = a_k(x_J^r)$;修改 M^r 的 $M(i,j)$和 $M(j,i)$同时修改 MAS_i,MOS,NS_i 转步骤 2;

(ii)否则,如果存在 $j_0 \in NS_i^r$,满足 $a_k(x_{j0}^r) \neq *$,则 $a_k(x_i^{r+1}) = a_k(x_{j0}^r)$;修改 M^r 的 $M(i,j)$和 $M(j,i)$同时修改 MAS_i,MOS,NS_i 转步骤 2;

步骤 3 删除噪声数据,算法结束。

3.2 改进后的算法分析

在填补能力上,改进后的算法根据辩识矩阵的已知信息,在无差别对象中,分析一行数据中属性 k 的出现次数,出现次数越少,则这个属性在遗失属性中出现的概率越大。扩充的辨识矩阵中,生成规则是 $a_k(x_i) \neq a_k(x_j) \cap a_k(x_i) \neq * \cap a_k(x_j) \neq *$,也就是说如果对象 i,j 的 k 属性不相等,则 $M(i,j)$中属性 k 是不会出现的,如果 M 的 i 行中属性 k 出现的次数也多,就说明对象 i 的 k 属性值出现的频率很低,反之该行中属性 k 出现的次数越少,则对象 i 的 k 属性值出现的频率越高,用出现概率大的属性填充,从而保证得到信息系统有尽可能高的支持度,改进后的算法能够填充原算法需要借助平均值等其它算法才能填充的属性,从而填充的结果能达到满意的程度,扩充了原算法的填补能力。

从时间复杂度上分析,改进后的算法时间复杂度是 $O(n*k)$,整个算法不需要原算法中的循环迭代,在填充遗失属性时,只修改 M,MOS,MAS_i,NS_i 中个别的值不重新计算新的可辨识矩阵,时间性能上有较大的改进。同时算法对辨识矩阵进行了化简,减少了运算量,且算法因为不需要每次判断新信息系统与旧信息系统的差异,所以没有外层的大循环,这在时间性能上也有很大的改进。

从区分噪声数据的能力上分析,改进后的算法能够区分出不完备的噪声数据,算法结束后有空缺的数据一定是噪声数据。在这里给出简单的证明(一个属性列全为空的信息系统除外)。假如这条数据不是噪声数据 a,也就是说,肯定存在一条完备的数据或者不完备数据(但与假定的数据 a 不同是一个遗

失属性)b 与假定的数据 a 存在相似关系，即肯定满足算法规定的一种填充标准,至少,这个属性值应该为 b 的属性值,所以算法最后输出是完备的信息系统，但最后输出是不完备的,矛盾,故得证。

从填充的效果上分析,由于改进后的算法取填充值是从相似对象中取的,这就去除了原算法步骤 3 可能是填充后的信息系统产生噪声数据的可能,同时还保证了原算法的填充效果。

4　实验结果

4.1　填补能力和不完备噪声数据的区分

设有一个不完备的信息系统 $S=(U,A,V,f)$,以表 2 中的数据作为算法的输入。

表 2　原始数据

U	a	b	c	d
1	no	yes	*	yes
2	*	yes	no	*
3	yes	no	yes	*
4	no	*	no	yes
5	yes	*	no	no
6	yes	yes	*	no
7	yes	*	*	*

经过两个算法的运算，最后得到两个新的信息系统 $S^0=(U^0,A,V,f^0)$，表 3 给出的是 ROUSTIDA 算法的输出结果,$S^1=(U^1,A,V,f^1)$。

从表 3 中可以看出第 2 条和第 7 条数据没有填补上,这需要执行 ROUSTIDA 的第 4 步算法，也就是借助于其他填充算法来填充。

从表 4 中看出,第 2 条和第 7 条数据所有的遗失属性都填补上，这说明了改进后的算法在填充能力上比原算法要优越;虽然表 2 中的第 3 条数据的遗失属性也被填充上了,但这条数据是噪声数据,没有一条规则与之相同,而表 4 却没有填补上,也就是说,改进后的算法能够区分出这条噪声数据。

表 3　ROUSTIDA 的输出

U	a	b	c	d
1	no	yes	no	yes
2	*	yes	no	*
3	yes	no	yes	no
4	no	yes	no	yes
5	yes	yes	no	no
6	yes	yes	no	no
7	yes	*	*	no

表 4　改进后的 ROUSTIDA 输出

U	a	b	c	d
1	no	yes	no	yes
2	yes	yes	no	no
3	yes	no	yes	*
4	no	yes	no	yes
5	yes	yes	no	no
6	yes	yes	no	no
7	yes	yes	no	no

4.2　时间复杂度和运行时间

实验中，对 100 条、200 条、300 条、400 条、500 条、600 条数据分别用两种算法进行填充，图 1 给出每种数据量下的 10 次实验的平均运行时间,可以看出,改进后的算法比原算法在运算时间上有了很大的改进。

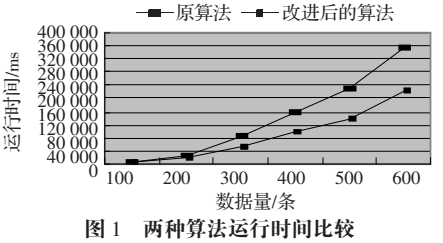

图 1　两种算法运行时间比较

5　结论

本文通过对基于粗糙集理论的不完备数据填补方法的改进,从而使其在遗失数据填充能力上有很大的改善,同时还具备了基本的噪声数据分离的功能,在时间性能上也比原来的算法有所改进,这对随后的高支持度规则的挖掘,以及整个系统性能的改善有一定的前导意义。实验表明,这是一种有效的填补方法,在现实的数据预处理系统中具有一定的应用价值。

参考文献:

[1] Kryszkiewicz M.Rough set approach to incomplete information systems[J].Information Science,1998,112(4):39-49.

[2] 王国胤.Rough 集理论与知识获取[M].西安:西安交通大学出版社,2001:17-19,41-42,96-99.

[3] 鄂旭,高学东,武森.一种新的遗失数据填补方法[J].计算机工程,2005,31(20):6-7.

[4] Zhu Wei-hua,Zhang Wei,Fu Yun-qing.An incomplete data analysis approach using rough set theory[C]//Proceedings 2004 International Conference on Intelligent Mechatronics and Automation 2004.New York:IEEE,August 2004:332-338.

[5] Kohavi R,Frasca B.Useful feature subsets and rough set reducts[C]//Proc on RSSCP'94,Proceedings of 3rd International Workshop on Rough Sets and Soft Computing,NY:pri Murt,USA,1994:200-244.

(上接 174 页)

4　结论

本文提出由有限论域上的 Vague 集导出的分拆真模糊向量和分拆非假模糊向量的定义,进而提出真近似推理模型和非假近似推理模型的定义及求法,使得模糊近似推理有了推广到 Vague 集近似推理的途径。实例表明这种基于模糊近似推理的 Vague 集双向近似推理的方法是可行的和有效的。

参考文献:

[1] Zadeh L A.Fuzzy sets[J].Information and Control,1965,8(3):338-353.

[2] Gau W L,Buehrer D J.Vague sets[J].IEEE Transactions on Systems,Man,and Cybernetics,1993,23(2):610-614.

[3] Bustince H,Burillo P.Vague sets are intuitionistic fuzzy sets[J].Fuzzy Sets and Systems,1996,79:403-405.

[4] 何平,王鸿绪.模糊控制器的设计及应用[M].北京:科学出版社,1997.

[5] 李凡.一个新的基于 Vague 集的近似推理方法[J].华中理工大学学报,2000,28(9):16-17.

[6] 李凡.一个基于 Vague 集相似度量的近似推理方法[J].计算机工程与科学,2002,24(5):98-101.

[7] 李凡.基于 Vague 集的元素间相似度量的近似推理[J].应用科学学报,2002,20(2):178-182.

[8] 王天江.基于 Vague 集的双向近似推理[J].华中科技大学学报:自然科学版,2002,30(8):21-23.

[9] 卢正鼎,王天江,李凡.基于 Vague 聚类的双向近似推理[J].小型微型计算机系统,2003,24(11):1933-1937.

[10] 李凡.基于 Vague 集的近似推理方法[J].华中科技大学学报:自然科学版,2004,32(4):44-46.

[11] 王天江. 一个新的基于 Vague 集的加权相似度量的双向近似推理方法[J].小型微型计算机系统,2004,25(2):211-215.

[12] Chun M G.A similarity-based bidirectional approximate reasoning method for decision-making system[J].Fuzzy Sets and Systems,2001,117:269-278.