文章编号:1000 - 582X(2002)03 - 0018 - 04

基于粗集理论的数据离散化新算法*

赵 军[1,2],王国胤[2],吴中福[1],李 华[1]
(1. 重庆大学 计算机学院,重庆 400044; 2. 重庆邮电学院 计算机科学与技术研究所,重庆 400065)

摘 要:连续属性值的离散化一直是机器学习领域中亟待解决的关键问题之一,他对于提高后继学习算法的运行速度、降低算法的实际空间要求和时间消耗、提高学习结果的聚类能力等都具有极其重要的意义。本文首先分析了基于粗集模型的数据离散化方法的特点和基本思路,研究了候选断点重要性的衡量方式,在此基础上提出两种新的从候选集合中最终确定离散化断点的启发式算法。这两种算法考虑并体现了粗集理论的基本特点和优点,选择的断点都能够保证信息系统的分辨关系,并能够取得较理想的离散化结果。

关键词:粗集;分辨关系;离散化;断点
中图分类号:TP18 **文献标识码**:A

决策系统中连续属性的离散化,即实型属性空间向整型属性空间的映射问题对数据挖掘而言具有非常重要的意义:首先,若决策系统中存在连续属性,那么对一个新的待识别样本而言,将很难通过其属性值矢量与训练数据表进行数据匹配的方式而得到识别,而有效的离散化就会显著地提高系统的聚类能力,增强系统对输入的待识别样本中数据噪音的鲁棒性;其次,离散化结果将会减小系统对存储空间的实际需求、加快后继数据挖掘算法的运行速度,减小后继算法的空间开销;此外,若离散化过程将某一连续属性的所有属性值均映射到同一结果,则该属性存在与否都不会影响系统对样本的分辨能力,因而它可以被删除。从这一角度来说,离散化过程同时也是属性约简的过程。

1 基于粗集理论的数据离散化

在处理不确定、不精确的含糊信息方面,粗集理论[1,2]具有不需要外界信息或先验知识的独特优点[3,4],人们对这一工具在数据挖掘领域的应用越来越重视,对基于粗集模型的离散化方法也进行了相应的研究,并取得了一些有价值的研究成果。这些成果大致可以分为两类:其中一类方法很少甚至是完全不考虑粗集理论的特殊性,直接把其它学科中的相应方法用来对信息系统进行离散化,其结果往往不能保证原有信息系统的分辨关系;另一类方法则充分考虑了粗集理论的特殊要求,其结果能够保证信息系统的分辨关系。其中,前一类方法的典型代表是粗集工具软件 Rosetta 提供的离散化方法**;后一类方法的典型代表是基于布尔逻辑和粗集理论的离散化方法[5,6]。

数据的离散化在其它领域也被称为"量化",它并不是一类新课题,人们已经对它进行了大量的深入研究,取得了丰硕的研究成果。但是,离散化问题也不是各学科可以完全通用的研究课题,实际上它在不同领域中有自己独特的要求和处理方式。由于粗集理论最核心的概念就是样本之间的"分辨关系"[3,4],因此,基于粗集理论的数据离散化方法要求能够保持信息系统所表达的样本分辨关系,否则将会导致信息丢失或者引入错误信息,从而影响所得结果的准确性。

对连续属性的离散化过程,从本质上来看,就是用一定的阈值(粗集的离散化理论中称之为"断点")对属性空间进行划分的过程。为了提高后继阶段的机器学习结果的聚类能力和识别能力,离散化过程要求防止对属性空间的过分细化。在保证离散化结果性能要

* 收稿日期:2001-10-16
基金项目:国家自然科学基金(69803014)、攀登特别支持费、重庆市科委攻关基金资助
作者简介:赵军(1971 -),男,重庆人,重庆大学博士研究生。主要研究领域为智能数据分析和处理、计算机网络与通信、现代远程教育技术等。

** Knowledge Systems Group. Rosseta Technical Reference Manual, 1999

求的前提下，用尽可能少的断点将属性空间划分成尽可能少的子空间，就成了离散化算法的追求目标。文献[5]认为：在保证信息系统分辨关系的前提下，采用基数最小的断点集合对系统进行的离散化就是基于粗集理论的最优离散化。从这一定义不难发现，对一个给定的信息系统，存在一种或多种最优的离散化结果。但遗憾的是，人们已经证明连续属性的最优离散化问题是一个NP难题[5,7]，因此对具有丰富样本的信息系统而言，求得最优离散化结果的时间开销将是令人无法忍受的，于是研究人员转而研究并提出一些实际可行的启发式算法，试图通过这些启发式算法获得次最优的离散化结果[5,6,8,9]。

基于粗集理论的数据离散化工作一般分3步进行：1）确定用于对各连续属性进行离散化的候选断点（阈值）集合；2）根据一定的策略选取候选断点集合的一个尽可能小的子集作为离散化过程实际采用的断点集合，即确定结果断点子集；3）利用结果断点集合对信息系统进行离散化。其中前两步是关键，粗集理论的特点要求这两步都必须保证系统对决策属性的分辨关系。为了从候选集合中选择适当的断点，需要采取有效的方式来度量和区分各候选断点的重要性。文献[5]为结果断点集合的确定提供了一种思路。这种方法首先求出信息系统中所有可能的连续属性断点值并用符号来标记，如 P_j^i 表示连续属性 a_i 的第 j 个断点值，称之为"符号断点"或直接称为"断点"；然后根据原始信息表 A 构造新的信息表 A^*。A^* 表达与 A 相同的样本分辨关系，它以符号断点为列索引，以 A 中所有能够分辨（具有不同决策值）的样本对 (x_i, x_j) 为行索引，表中的元素 $A^*[(x_i, x_j), P_j^i]$ 这样来确定：若样本对 (x_i, x_j) 能够被断点 P_j^i 分辨，则对应的元素值为1（称为"1元素"），否则为0（称为"0元素"）；最后利用表 A^*，从候选断点集合中选择合适的子集来对信息系统进行离散化。

从思路上来说，基于表 A^* 的启发式算法可以分为"前向选择法"和"反向删除法"两类：前一类是首先将结果集合初始化为空集，然后逐步从表 A^* 中选取当前重要性程度最高的断点加入到结果集合中，直至结果集合中的断点已经能够保证信息系统的分辨关系为止；后一类是首先将结果集合初始化为全集，即包含所有可能的断点，然后再从结果集合中逐步去除当前重要性程度最低的断点，直至将要改变信息系统的分辨关系为止。对这两类方法而言，如何衡量断点的重要性都是至关重要的。由于对候选断点重要性的衡量方式及其结果是决定算法最终结果的一个关键因素，因此衡量标准的合理性将会直接影响到算法的性能。

表1中断点的重要性可以从两个方面来衡量：从列方向来看，某列值为1的元素个数越多，则对应的断点对样本的分辨能力越强，断点的重要性越高；从行方向来看，某行值为1的元素个数越少，则表明能够分辨对应样本的断点越少，相应地这些断点的重要程度越高。启发式算法在选择断点时，主要依据断点的这些特性。显然，列方向和行方向特征对断点重要性的反映方式是不同的，前者能够直接反映断点对样本的分辨能力，后者则需要考察该行对应的样本对才能获得相应的信息，是一种间接的方式。

2　基于粗集的数据离散化新算法

文献[5]、[6]在选择断点时，采用了"贪心算法"的思想，仅利用了断点列方向的特征。针对这一点，文献[8]则试图同时以列方向特征和行方向特征作为断点的选择依据，提出了两种改进算法。在具体实现方法上，这两种改进算法均考察了满足一定特征的数据行，统计这些数据行上值为1的元素个数之和，并分别以求得的和作为"改进算法一"选择断点的辅助依据及"改进算法二"选择断点的主要依据。已知，从不同的行来看，同一断点可能具有不同的重要程度，选择断点时，应该参考能够反映出断点最高重要性的行。但这两种改进算法选择断点的依据都涉及到统计多行"1元素"的总个数，而不是单独考虑各行，这种方式过于笼统，必然会掩盖那些具有特殊意义的行所表达的特征，因此这种依据本身具有一定的不合理性。比如若在表 A^* 中有两个断点 P_p^q 和 P_r^s。若前者能分辨的样本对中，对应行上"1元素"个数的最小值为2，即该样本对仅能被两个断点区分；而后者能够分辨的样本对中行上"1元素"的最少个数为3，即该样本对能被3个断点区分。为了区分相应的样本对，选择 P_p^q 的机率是50%，而选择 P_r^s 的机率只有33%。采用多行求和的方式将不能提供这些信息，在某些情况下不能优先选择机率更高的断点。从文献[8]给出的仿真结果来看，改进算法二以"多行和"作为选择断点的主要依据，其性能与改进算法一相比，存在较大的差距。从某种意义上来说，这样的结果反映出该算法所采用的衡量断点重要性的方式具有较强的局限性。

在此提出两种新的算法。一种方法在选择断点时以列方向特征为主要依据，以行方向特征为辅助依据；与之相对应，另一种方法则主要考虑行方向特征，辅之以列方向特征。这两种算法的优点在于充分利用了表 A^* 提供的信息，既利用了信息表 A^* 列方向的特征，

也利
用了其行方向的特征;同时对行方向特征,我们根据单行的信息来确定断点的重要性,摒弃了文献[8]中多行和的方式。同时,这两种算法都能够保证原始信息系统的分辨关系。算法的形式化描述如下。

算法一(列先行后法)描述:

定义函数 SUM(Row):计算行上值为 1 的元素个数;SUM(Column):计算列上值为 1 的元素个数。

信息表 A^* 及初始化为空的断点集合 Π。

1) 对所有的列:计算 SUM(Column)

2) 逐一考察具有最大 SUM(Column) 的列 $C_{i1}, C_{i2}, \cdots, C_{im}$:

对所有的 $C_{it}(t \in ([1,m])$:求出列上元素值为 1 且对应行上 SUM(Row) 具有最小值的行(若有多行同时满足该条件,则取最先出现的行)所对应的列索引 P_j^a 加入集合 Π,并消去列 P_j^a 上所有值为 1 的元素所在的行及列 P_j^a 本身。

4) 若信息表 A^* 不为空则转步骤 1),否则转步骤 5)。

5) 算法终止。集合 Π 即为所求的断点集合。

算法二(行先列后法)描述:

同样定义函数 SUM(Row):计算行上值为 1 的元素个数;SUM(Column):计算列上值为 1 的元素个数。

信息表 A^* 及初始化为空的断点集合 Π。

1) 对所有的行:计算 SUM(Row);

2) 逐一考察具有最小 SUM(Row) 的行 $L_{i1}, L_{i2}, \cdots, L_{im}$;

对所有的 $L_{it}(t \in [1,n])$:求出行上元素值为 1 且对应列上 SUM(Column) 具有最大值的列,若有多列同时满足该条件,则取最先出现者。将该列的索引 P_j^a 加入断点集合 Π,并消去该列上所有值为 1 的元素所在的行及该列本身。

3) 若信息表 A^* 不为空则转步骤 1),否则转步骤 4);

4) 算法终止。集合 Π 即为所求的断点集合。

3 算法应用实例

考察如表 1 所示的信息表 A^*:

表 1 信息表 A^*

A^*	P_1^a	P_2^a	P_1^b	P_2^b	P_3^b
x_1, x_2	0	0	1	0	1
x_1, x_3	0	0	1	1	0
x_1, x_5	1	0	0	1	0
x_2, x_4	1	0	1	1	1
x_2, x_6	1	1	0	0	0
x_2, x_7	1	1	0	0	0
x_3, x_4	1	1	0	0	0
x_3, x_6	1	1	0	0	0
x_3, x_7	0	1	1	0	0
x_4, x_5	0	1	1	0	0

对表 1 所示的信息系统,按文献[8]的改进算法一,得到的断点集合将包含 3 个断点;按文献[8]的改进算法二,得到的断点集合将包含 4 个断点。而按本文提出的两种算法,都将会得到断点集合 $\{P_1^a, P_1^b\}$,这明显优于文献[8]的结果。容易证明,按文献[5]中的相应定义,这一结果是本信息系统的一个最优断点集合。

为进一步验证算法的有效性和比较算法的性能,笔者对 UCI 机器学习数据库中的一些实际数据集进行了分类规则知识获取实验:随机将数据集分为相等的两个部分,一半数据作为训练样本,另一半数据作为测试样本。对训练样本采用不同的方法对原始数据进行离散化处理,然后对离散化结果进行同样的处理:首先进行属性约简,再运用 Skowron 算法获取决策规则,阈值均取 0.6,最后用测试样本对得到的规则进行测试。由于启发式算法运行的结果与数据分布等诸多因素有关,为尽可能客观地评价算法的性能,笔者对每一数据集均进行了 5 次随机数据划分,对每一划分都进行了上述实验。表 2、3 给出了几个指标的平均结果。

从仿真结果来看,与参考离散化算法相比,本文中提出的第 1 种(列先行后)算法在断点数、误识率和拒识率等性能指标上都能够得到较好的结果,算法的综合性能优于参考算法,而第 2 种(行先列后)算法的性能则相对较差。

表 2 算法断点选取平均结果对比表

数据集合	Nr. of Real Attr.	Nr. of Objects	贪心算法		文[8]改进算法 1		本文提出算法 1		本文提出算法 2	
			A	*B*	*A*	*B*	*A*	*B*	*A*	*B*
Ecoli	7	336	5.0	14.2	5.2	14.4	5.2	12.2	5.2	11.8
Glass	9	214	7.6	11.2	7.7	11.6	7.2	9.6	7.4	10.6
Iris	4	150	2.6	4.4	2.4	4.4	2.4	4.4	3.2	6.0

注:表 2 中符号"*A*"、"*B*"分别指剩余连续属性数目和最终选取的断点数目。

表 3　算法测试平均结果对比表

数据	贪心算法			文[8]改进算法 1			本文提出算法 1			本文提出算法 2		
集合	I	II	III	I	II	III	I	II	III	I	II	III
Ecoli	73.0	26.8	0.2	72.8	27.2	0.0	75.4	24.4	0.2	72.2	27.6	0.2
Glass	52.0	48.0	0.0	55.0	45.0	0.0	57.2	42.8	0.0	49.2	50.8	0.0
Iris	95.0	5.0	0.0	95.4	4.6	0.0	95.0	5.0	0.0	91.4	8.6	0.0

注：表 3 中符号"I"、"II" 和"III" 分别指识别率、误识率和拒识率。

4　结论

在按文献[5]构造的信息表 A^* 中，可以独立地根据表的各行和各列上的特征来衡量属性断点值的重要性。根据这一重要结论，笔者提出了两种数据离散化算法。这两种算法都能够保证信息系统原有的分辨关系，并且仿真结果表明，由于"列先行后" 算法合理地、充分地考虑和利用了信息表 A^* 行方向和列方向的特征，其综合性能明显优于参考算法。

同时，算法的仿真结果表明，表 A^* 行方向特征对断点重要性的表征不如列方向特征准确。从结果断点的个数、规则的正确识别率和拒绝识别率等指标来看，已经提出的各种以行方向特征为主的离散化算法的性能都不如对应的以列方向特征为主的算法，其原因在于列方向特征能够直接反映断点对样本的分辨能力，而行方向特征则需要通过考察对应的样本对来获得相应的信息，是一种间接的方式。

参考文献：

[1] PAWLAK Z. Rough Set[J]. Int. J. of Computer and Information Science, 1982, 11: 341 - 456.

[2] PAWLAK Z. Rough Set[J]. Communications of the ACM, 1995, 38(11):89 - 95.

[3] 曾黄麟. 粗集理论及其应用 - 关于数据推理的新方法(修订版)[M]. 重庆：重庆大学出版社，1998.

[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，2001.

[5] NGUYEN HS, SKOWRON A. Quantization of real values attributes, rough set and boolean reasoning approaches[A]. Proc. of the 2nd Joint Annual Conf. on Information Sci.[C]. USA, Wrightsville Beach, NC, 1995,34 - 37.

[6] NGUYEN H S. Discretization problem for rough sets methods[A]. Proc. of the 1st Int. Conf. on Rough Sets and Current Trends in Computing(RSCTC'98)[C], Warsaw, Poland, 1998,545 - 552.

[7] 权光日，文光远，叶风，等. 连续属性空间上的规则学习算法[J]. 软件学报，1999, 10(11): 1 225 - 1 232.

[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] NGUYEN SH . Some efficient algorithms for Rough Set methods [A]. Proc. Of the Conf. Of Information Processing and Management of Uncertainty in Knowledge Based Systems[C], Granada, Spain, 1996: 1 451 - 1 456.

New Algorithms for Data Discretization Based on Rough Set Theory

ZHAO Jun[1,2], *WANG Guo-yin*[2], *WU Zhong-fu*[1], *LI Hua*[1]

(1. College of Computer Science&Engineering of Chongqing University, Chongqing 400044, China;
2. Institute of Computer Science&Technology of Chongqing Univ. of Posts&Telecomms, Chongqing 400065, China)

Abstract: The discretization of real values is always one of the key problems to be solved in the domain of machine learning for its great contribution to speeding up the followed learning algorithms, cutting down the real demand of algorithms on running space and time, and improving the clustering capability of the ultimate learning results. The basic characteristics and framework of discretization approaches based on rough set model are analyzed at first, then the different measurements of the importance of candidate cuts are discussed and researched. Two new heuristic algorithms are put forward to finally select the useful cuts from a candidate set. The selected cuts of the two algorithms will adequately maintain the discernible relation of information systems for their full considering the specialty of rough set, which perfectly embodies the advantages of this theory. Moreover, excellent discretization results may be expected through these heuristic algorithms.

Key words: rough set; discernible relationship; discretization, cut

(责任编辑　吕赛英)