基于分形维数的属性约简

郭 平 陈其鑫 王艳霞
(重庆大学计算机学院 重庆 400044)

摘 要 关于属性约简的算法已经提出了许多,基于粗糙集的属性约简算法就是其中的一类。但该类算法执行效率低且不一定得到最小约简。本文讨论了基于可辨识矩阵的属性频度算法(BDMF)并提出了基于分形维数的向后剔除属性约简算法(FDR)。仿真实验表明 FDR 比 BDMF 的运行效率高,且约简的效果更好。

关键词 属性约简,分形维数,可辨识矩阵,属性频度,粗糙集

Attribute Rdeuction Basde on Fractal Dimension

GUO Ping CHEN Qi-Xin WANG Yan-Xia
(School of Computer Science, Chongqing University, Chongqing 400044)

Abstract Among those algorithms of attribute reduction proposed, some based on rough set. However, this type of algorithm is not efficient enough and also minimum reduction would not necessarily achieved by them. In this pqper, the algofithm on attribute frequency based on identification matrix algorithm (BDMF) is discussed and then the algorithm based on Fractal Dimension Reduction algorighm (FDR) was developed. It is shown that FDR has higher running effciency and more effective reduction than BDMF.

Keywords Attribute reduction, Fractal dimension, Discrimination matrix, Attribute frequency, Rough set

1 引言

在数据挖掘、文档分类和多媒体索引等领域中,所面临的数据对象往往是大数据集,其中包含的属性个数和记录个数都很大(如,几十个甚至上百个属性),由此导致处理算法的执行效率低下。众所周知,数据集中的属性并不是同等重要的,甚至其中某些属性是冗余的。特别是随机采集数据的冗余度更大。冗余数据的存在,一方面是对资源(存储空间)的浪费;另一方面,干扰人们利用这些数据做出正确的决策。属性约简就是在保持数据分类或决策能力不变的条件下,删除其中冗余的或不重要的属性[2]。

粗糙集理论中用决策表来表示含有决策属性的数据集,基于决策表的属性约简是粗糙集理论的核心内容之一[1]。决策表可能存在多个属性约简,因此,找到具有最少属性的约简(即最小约简)一直是研究者追求的目标。文[5,6]中提出了基于神经网络的约简算法,文[7]给出了基于遗传算法的约简算法,基于粗糙集的约简算法在文[8]中被讨论了。然而,S. Wong 和 W. Ziarko 在 1985 年就已经证明:基于粗糙集的决策表最小约简是 NP-hard 问题[3]。导致属性约简是 NP-hard 问题的主要原因是属性组合爆炸。利用与数据集的特征,在算法中融入启发性知识,缩小问题求解的搜索空间则是一类被认为是更有意义的算法[4]。

尽管以上各种算法都提出了进行属性约简的详细步骤和方法,但是都无法说明得到的约简是否为最小约简。基于粗糙集的属性约简算法是目前流行的方法,但同样不一定得到最小约简。基于分形维数的属性约简算法[8]可以有效地改变这一状况。

本文利用数据集的分形维数进行属性约简。第 2 部分介绍一些基本概念,第 3 部分分别讨论了基于可辨识矩阵的属性频度算法(BDMF)和基于分形维数的向后剔除属性约简算法(FDR),最后给出了 BDMF 和 FDR 的实验比较,最后总结了全文。

2 基本概念

2.1 分形和分形维

如果一个数据集在所有的观察尺度下都具有自相似性,即一个数据集的部分分布有着与整体分布相似的结构或特征,称该数据集是分形的[9]。下面给出关于数据集维数的几个概念。

定义 1 嵌入维(Embedding dimension) 数据集中的数据点所在欧式空间的维数称为数据集的嵌入维,即一个数据集中属性的个数。

定义 2 固有维(Intrinsic dimension) 一个数据集的固有维是指一个数据集所表示的空间对象的实际维数。

一般地说,空间对象的维数(固有维)不会超过所在欧式空间的维数(嵌入维)。例如,所有欧氏空间的直线不论嵌入维是二维还是三维,其固有维都是一维的。

定义 3 分形维(Fractal dimension) 嵌入维数等于 n 的数据集可视为 n 维空间中的点。用边长为 r ($r\in(r1, r2)$)的 n_维立方体分割数据集,记落入第 i 个立方体中的数据点的数目为 C_i。则分形维 D_q 计算如下:

$$D_q=\frac{1}{q-1}\times\frac{\partial\log\sum_i D_i^q}{\partial\log r}\quad r\in(r1, r2) \tag{1}$$

其中,D_0 称为 Hausdorff 分形维;当 q 趋近于 1 时,D_1 称为信息分形维(Information Fractal dimension);当 $q=2$ 时,D_2 称为相关分形维(Correlation fractal dimension)。D_2 描述了随机选择的两个数据点的距离落在某一范围内的概率,因此相关分形维 D_2 的改变意味着数据集中数据点分布的变化。

下面使用相关分形维作为数据集固有维的度量。

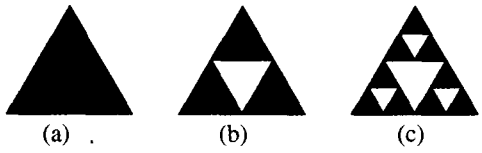

图 1 Sierpinsky 三角形的递归机构

例如，图 1(a)是边长为 1 的等边三角形。从图 1(a)中去掉中间的等边三角形得图 1(b)(阴影部分)；再从图 1(b)中去掉每个三角中间的等边三角形得图 1(c)。如此继续下去，当进行无穷步后将剩下的三角形的集合称为 Sierpinsky 三角形。显然，Sierpinsky 的嵌入维数等于 2，但它的相关分形维是 $\log(3)/\log(2)=1.58$[11]。

在实际应用中，计算数据集的相关分形维几乎是不可能的，计盒(box-counting)维数常被用来近似估计相关分形维数。计盒维数可按以下方法计算[8]：

假设数据集中有 N 个数据点，每个数据点具有 E 个属性，则这 N 个点可看作为 E 维空间中的点$(x_1x_2\cdots x_E)$。将每维的取值范围 $1/r$ 等分($r=1/2,1/4,1/8,\cdots$)，由此数据空间被划分为 E 维网格，整个空间被分成$(1/r)^E$ 个单元格。从 0 开始依次为每个单元格编号，记为 $0,1,2,\cdots,(1/r)^E-1$，其中编号为 0 的单元格的点的坐标范围为$(0,0,\cdots,0,0)$到$(0,0,\cdots,0,R)$，R 为单元格的半径；编号为$(1/r)^E-1$ 的单元格的点的坐标范围为$(R,R,\cdots,R,0)$到$(R,R,\cdots,R,R)$。令

$$T_1=\text{int}(x_{i1}/R),\ T_2=\text{int}(x_{i2}/R),\cdots,\ T_j=int(x_{ij}/R),\cdots,\ T_E=\text{int}(x_{iE}/R)$$

则第 i 个点$(x_{i1}x_{i2}\cdots x_{iE})$在编号为 $T_1\times(1/r)^0+T_2\times(1/r)^1+\cdots+T_j\times(1/r)^{j-1}+\cdots+T_E\times(1/r)^{E-1}$ 的单元格内。用 $C_{r,i}$ 表示 $1/r$ 等分时落入第 i 个单元格中的数据点数。令

$$S(r)=\sum(C_r^2,\ i)$$

则数据集的分形维数(计盒维数)为：

$$-\frac{\partial\log s(r)}{\partial\log r} \tag{2}$$

对于具有自相似特征的数据集，(2)式的值为常数。对于实际数据集，以 $\log r$ 为横坐标，$\log S(r)$ 为纵坐标的曲线中近似为直线部分的斜率常被视为数据集的分形维的近似值。

2.2 可辨识矩阵

定义 4 设 C、D 分别是数据集 $U=\{x_1,x_2,\cdots,x_n\}$的属性集，f 是 U 到 V 的映射。称 $S=(U,A,V,f)$是一个知识表达系统，其中：$A=C\cup D$，$C\cap D=\Phi$，C 称为条件属性集，D 称为决策属性集。

特别，当 $C\neq\Phi$ 且 $D\neq\Phi$ 时，知识表达系统称为决策表。

定义 5 给定一个决策表 $S=(U,A,V,f)$，$A=C\cup D$ 是属性集，C、D 分别是条件属性集和决策属性集。可辨识矩阵[12] $M=(c_{ij})$ 定义为：

$$(c_{ij})_{n\times m}=\begin{cases}\{a\in C|a(x_i)\neq a(x_j)\} & D(x_i)\neq D(x_j)\\ 0 & D(x_i)=D(x_j)\\ -1 & a(x_i)\neq a(x_j)\text{ 且 }D(x_i)\neq D(x_j)\end{cases} \tag{3}$$

其中 $a(x)$ 是元组 x 在属性 a 上的取值，$D(x)$ 是 x 在决策属性 D 上的取值。

3 属性约简算法

本节我们将讨论两个属性约简算法，在给出算法的同时给出算法的计算复杂度分析。

3.1 基于可辨识矩阵的属性约简算法(BDMF)

在信息系统中，可辨识矩阵中的元素值代表了可辨识记录的条件属性集合。在可辨识矩阵中，出现频率高的属性表示该属性可辨识的记录较多，而出现频率低的属性表示该属性可辨识的记录较少。极端的情况是在可辨识矩阵未出现的属性表示不能用于区分记录可以直接删除。因此，我们可以将属性在可辨识矩阵中出现次数的多少作为属性重要性的判断依据：若某属性在可辨识矩阵中出现的次数越多，表明该属性可辨识能力越大，其重要性越高；反之，若某属性出现的次数越少，表明该属性可辨识能力越小，其重要性越低。由此，根据属性在可变辨识矩阵中出现的频率获得属性约简算法 BDMF。

算法：BDMF

输入：决策表 $S=(U,A,V,f)$，$A=C\cup D$，$C\cap D=\Phi$，C 为条件属性集合，D 为决策属性集合；

输出：约简后的属性集 Redu；

步骤：

第 1 步：计算决策表的可辨识矩阵 M。将可辨识矩阵中的核属性 C_0(即属性组合数为 1 的条件属性)赋给属性约简后得到的属性集，即 Redu$\leftarrow C_0$；

第 2 步：去掉可辨识矩阵中含有核属性的属性组合项；

第 3 步：计算可辨识矩阵中所有剩余属性项中各条件属性出现的频率，记出现频率最高的属性为 a。计算 Redu$\leftarrow$Redu$\cup\{a\}$，将可辨识矩阵中包含有条件属性 a 的属性组合项删除掉；

第 4 步：若可辨识矩阵不为空则转第 3 步；

第 5 步：输出 Redu。

显然，算法 4 求取的是所有约简结果中的某一个或某一些(当某两个条件属性出现频率相同时)结果，当信息系统的复杂程度较高时，其求解的复杂度大大小于原来的约简方法。当然，该方法并不能保证在任何情况下求得的约简结果都是原信息系统的最小约简，但在它使得求解的复杂程度大大降低的情况下，是一种简单而有效的方法，在大多数情况下可以获得原信息系统的最小约简。

由(3)式，可辨识矩阵 M 具有对称性，在第 1 步中计算 M 时只需计算 $|U|(|U|-1)/2$ 项，因此计算可辨识矩阵的代价是 $O(|A|\ |U|^2)$。循环计算中(第 3 步)，可辨识矩阵最多有 $|U|(|U|-1)/2$ 项，每项最多包含 $|A|$ 个属性，最坏的计算代价为 $O(|A|\ |U|^2)$，所以 BDMF 算法的时间复杂度为：

$$T(BDMF)=O(|A|\ |U|^2) \tag{4}$$

3.2 后向剔除属性约简算法 FDR

决策表 $S=(U,A,V,f)$ 中的对象 $U=\{x_1,x_2,\cdots,x_n\}$ 显然可以看作 $E=|A|$ 维空间中的点集，即 S 中可用于区分数据对象的属性数目不超过 E。由此，可以通过计算 E 维空间中点集$\{x_1,x_2,\cdots,x_n\}$的分形维来估计 S 应包含的最少属性数。我们通过两个计算步骤来约简 S 的属性集：(1)计算包括所有 E 个属性的分形维，称为全分形维 wfd；(2)剔除其中的一个属性，再计算剩余的 $|A|-1$ 个属性的分形维，称它们为部分分形维 pfd，这样共计算出 E 个部分分形维：pfd_1，$\cdots$，$pfd_{|A|}$，从 $pfd_1,\cdots,pfd_{|A|}$ 中选择最接近于 wfd 的一

(下转第 239 页)

进一步研究新的、效果更好的手写体数字的特征提取方法具有重大的意义。

参 考 文 献

1 Chien-cheng , Tang Yun-ching. To improve the training time of BP neural networks. Info-tech and Info-net, 2001 International Conferences on, 2001, 3: 473～479
2 Hornik K. Approximation capabilities of multilayer feedforward networks. IEEE Transactions on Neural Networks, 1991, 4: 251～257
3 Yu X H, Chen G A. Efficient estimation of dynamically optimal learning. In: Proc. of IEEE ICNN-95, 1995. 385～388
4 Sakaue S, Kohda T, Yamamoto H, Maruno S, Shimeki Y. Reduction of required precision bits for Back-Propagation applied to pattern recognition. IEEE Transactions on Neural Networks, 1993, 4 (2): 270～275
5 Kwon T M, Chen Hui. Contrast enhancement for backpropagation. IEEE Transactions on Neural Networks, 1996, 7(2): 515～524
6 Chen J Q, Jiang J P. New method to train a BP network and their application. International Joint Conference on Neural Networks, 1999, 3
7 朱学芳. 手写数字识别实验系统的研究[J]. 南京大学学报, 1996. 1
8 谢光毅, 钟义信. 神经网络用于手写体数字识别[J]. 模式识别与人工智能, 1994, 12(4)
9 刘滨. 基于神经网络的车牌字符识别研究[D]: [武汉大学硕士学位论文]. 2004. 8
10 胡小锋, 赵辉. Visual C++/MATLAB 图像处理与识别实用案例精选[M]. 北京: 人民邮电出版社, 2004. 9
11 朱小波. 基于神经网络的手写体数字识别分析与研究[D]: [武汉科技大学硕士学位论文]
12 张捷. 手写数字识别的研究与应用[D]: [西安建筑科技大学硕士学位论文]

(上接第 190 页)

个 pfd_j，将其对应的属性 a_j 剔除，同时将 pfd_j 作为下一步的 wfd，在属性集 $A-\{a_j\}$ 中继续上述的步骤，直到剩余属性数目与 S 的分形维数相同。

我们获得后向剔除属性约简算法 FDR 如下。

算法：FDR

输入：决策表 $S=(U,A,V,f)$，$A=C\cup D$，$C\cap D=\Phi$，C 为条件属性集合，D 为决策属性集合

输出：最小约简 Redu

步骤：

第 1 步：算出决策表 S 的计盒维数 wfd（全分形维）；$wfd_0 \leftarrow wfd$；

第 2 步：For $1\leqslant i\leqslant |A|$ ($a_i\in A$) 计算 U 在 $A-\{a_i\}$ 属性集上的部分分形维 pfd_i；

第 3 步：从 $pfd_1,\cdots,pfd_{|A|}$ 中选择最接近于 wfd 的一个 pfd，记为；计算 $A\leftarrow A-\{a_j\}$；$wfd\leftarrow pfd_j$；

第 4 步：如果 $|A|>wfd_0+1$，转第 2 步；

第 5 步：输出 A。

在 FDR 中，计算有 N 个元组的决策表的分形维的时间复杂度为 $O(N^2)$；从 $|A|$ 个属性中选择一个属性并剔除需要扫描 $|A|$ 次对象集 U，如果剔除 K（$K<|A|$）个属性，则需要扫描 $(K\times(2|A|-K+1))/2$ 次，所以整个算法的时间复杂度为 $O(|A|\times K\times N^2)$。

4 算法比较分析

为比较算法 BDMF 与算法 FDR 的效率，我们在 Intel 2.8GHz CPU×2（内存 1024M）上对具有 6 个属性（a_1,a_2,a_3,a_4,a_5,a_6）5500 条记录的合成数据集 FHMIN Test 6 进行了实验对比。其中，属性 a_1,a_2,a_3 对应的值由随机函数生成，属性 a_4,a_5,a_6 的值分别计算为：

$a_4=\sin(x1+x2+x3)+3$，$a_5=\sin(x1*x2+x3)$，$a_6=|x2+x3*\sin(x1+x3)|$

实验的结果如表 1。

表 1 BDMF 与 FDR 对比实验

Record	BDMF		FDR	
	Running Time (s)	Reduction	Running Time(s)	Reduction
2000	25	[x1,y1,y2,x2,]	4	[x1,x2,x3,]
3000	53	[x1,x2,x3,]	11	[x1,x2,x3,]
4000	96	[x1,x2,x3,]	16	[x1,x2,x3,]
5500	297	[x1,x2,x3,]	32	[x1,x2,x3,]

实验结果分析：由于 BDMF 算法是基于可辨识矩阵的，故对于 m 个属性、n 个元组的决策表来说，它的可辨识矩阵就是一个 $\max(m,n)\times\max(m,n)$ 的方阵，现实情况中一般属性的个数 m 都不太大，但是元组的数量 n 却是大得惊人，这时不但运算的时间很长，而且由于要占有大量的内存，使得 BDMF 算法的适用范围大大降低，实验中发现当元组数量在增加时，花费的时间急剧增加；而分形维的计算由于不需要生成可辨识矩阵，受元组数量的影响要小得多，都很快地得到了满意的结果。当 Record∈[3000, 5500]时，BDMF 得到的约简为最小约简，其运行时间从 53 秒增加到 297 秒，而 FDR 不但在数据集的记录数在各个阶段都能获得正确的结果，且花费的时间也是可以容忍的。

结论 一个数据集的分形维体现了数据集的固有特征。算出数据集的分形维数就可以确定该数据集最小约简的大小，基于这种思想，本文利用后向剔除属性的约简算法 FDR 来进行属性约简并用基于可辨识矩阵的属性约简算法 BDMF 来进行对比分析。算法复杂性分析的结果和实验的结果均说明 FDR 在时间和空间两方面都优于 BDMF。

参 考 文 献

1 Pawlak Z. Rough Sets-Theoretical Aspects of Reasoning about Data [J]. Kluwer Academic Publishers, 1991. 6～42
2 Miao Duoqian, Wang jue. An information-based algorithm for reduction of knowledge [J]. IEEE ICIPS'97. 1991. 1155～1158
3 Wong S, Ziarko W. On optimal decision rules in decision tables [J]. Bulletin of Polish Academy of Science. 1985, 33: 693～696
4 Miao Duoqian, Wang jue. Analysis on feature reduction strategies of rough set[J]. Jo-urnal of Computer Science and Technology, 1998, 13(2): 189～192
5 Scherf M, Brauer M. Improving RBF networks by the feature selection approach EUBAF-ES[J]. In: W. Gerstner. ed. Proc. 7th Intl Conf on Artificial Neural Networks. Lausanne, Switzerland: Springer, 1997. 391～396
6 Robert A, Stocker E. Classification and feature selection by a self-organizing neural networ-k[J]. In: Dorffner G, ed. Proc. of Int Conf on Artificial and Neural Networks, UK: Springer 1999. 651～660
7 Pernkopf F, O'Leary P. Feature selection for classification using genetic algorithms with a novel encoding[J]. In: W. Skarbek, ed. Proc. of Computer Analysis of Images and Patterms, Warschau, Poland, Springer, 2001. 161～168
8 Traina C, Traina A, Wu L, et al. Fast feature selection using fractal dimension[J]. In: C. Faloutsos, ed. Proc. of XV Brazilian Symposium on Databases, Paraila, Brazil: Springer, 2000. 78～90
9 Talavera L. Feature selection as a preprocessing setp for hierarchical clustering[J]. In: I. Bratko, ed. Proc. of the 16th Int Conf on Machine Learning. Bled, Slovenia: AAAI Press, 1999. 389～397
10 Grassberger P. Generalized, Dimensions of Strange Attractors [J]. Physics Letters, 1983, 97A: 227～230
11 Schroeder M. Fractals, chaos, power larws, 6ed [M]. New York: W. H. Freeman and Company, 1991
12 王国胤. Rough 集理论与知识获取[M]. 西安交通大学出版社, 2001. 32～67