近似约简算法研究*)

谢晓飞[1] 邵 斌[2] 张建宏[2]

(浙江医药高等专科学校 宁波 315192)[1] (湖州师范学院信息工程学院 湖州 313000)[2]

摘 要 信息系统属性的约简可以提高知识发现、机器学习等的精度和效率。本文提出了一种近似约简算法,该算法可使信息系统在基本保持原风格的情况下尽可能少地保留属性,为后期的系统处理节约了大量的处理时间。该算法的时间复杂度没有提高,约简后的属性大大减少。虽然原信息系统有一定的损失,但在一定的显著水平下是可以接受的。最后对一个有9个属性的信息系统进行了约简和近似约简的对比分析。

关键词 属性约简策略,区分矩阵,近似约简

The Study of Approximately Reduction Arithmetic

XIE Xiao-Fei[1] SHAO Bin[2] ZHANG Jian-Hong[2]

(Zhejiang Pharmaceutical College, Ningbo 315192)[1] (School of Information Engineering, Huzhou Teachers College, Huzhou 313000)[2]

Abstract The attributes reduction of information system can enhance accuracy and efficiency of knowledge discovery, machine learning, etc. This paper proposes a approximately reduction arithmetic, this arithmetic can retain the minimal attributes in the basic form of the information systems, as much as possible to reduce attributes, that can save the time of the system's upper disposal. The arithmetic time order of the complexity has not enhanced, but the reduced attributes greatly reduced. The original information system has the certain loss, but this is may accept under the certain remarkable level. And a system having nine attributes been carried on the reduction and the approximately reduction of contrast analysis.

Keywords Attributes reduction strategy, Discernibility matrix, Approximately reduction

1 引言

在粗糙集理论中,属性约简是一个非常重要的问题。该问题可以定义为在初始的 N 个属性中找到 M 个属性(M 小于等于 N)来描述原始的数据集,从而使分类差错概率最小。特别是在处理大量数据时,属性约简变得更加重要,因为许多学习算法在数据约简之前就可能因为属性太多要消耗太多的时间以至于丧失使用价值。另外,不相关的冗余属性可能会降低预测精度,并减弱归纳出来的知识模型的理解度。

为此,已有不少学者和研究人员针对属性约简做了大量工作,也提出了不少算法[3~8],但这些算法策略都存在计算复杂度大,计算效率低的问题。

定义1 如 $S=(U,A)$ 上一个信息系统,其中 U 是对象的非空集合,A 是属性的非空有限集合,对于每一个 $a\in A$,有 $a:U\to Va$,其中 Va 称为 a 的值域。

定义2(相似关系) $R\subseteq A$, $SIM(R)=\{(x,y)\in U\times U \mid (a\in R, a(x)=a(y))\}$。令 $S_R(x)$ 表示对象集 $\{y\in U \mid (x,y)\in SIM(R)\}$。

对于 R 而言,$S_R(x)$ 是与 x 不可区分的对象的最大集合。

定义3(约简) 对于 $x(x\in U)$,一个集合 $R\subseteq A$ 是信息系统 S 关于 x 的一个约简,若 $S_R(x)=S_A(x)$ 且对于 $\forall B\subset R$, $S_B(x)\neq S_A(x)$。

基于区分矩阵的属性约简策略是实现每一个对象对在所有属性上的两两比较,也就是说,区分矩阵 M 非常庞大,若原先信息表 S 是一个 n 行 m 列矩阵,其中 n 个对象,m 个属性,那么区分矩阵 M 就是 $n\times n$ 矩阵。当数据量很大时,对矩阵 M 进行操作运算量很大,而且对区分矩阵的运算需要每次都要遍历整个区分矩阵。能否找到一个新的办法,运算量小,但是也能实现基于区分矩阵属性约简策略的功能。秦中广博士在其博士论文中提出一种新型的属性约简策略 ARS[9],该策略把属性值的个数应用到属性约简上,大大降低进行属性约简的计算复杂度和空间复杂度,在约简大量数据更能显示其优越性。

以一个信息表为例,有23个对象,其中有1个属性 A,其各对象的属性值在 H,M,LO 中取,那么按照区分矩阵的思想是23个对象两两比较,比较这些对象对在属性 A 上的取值,若该对象对在属性 A 上取值相同则为0,若不同则把属性 A 写上去,到最后计算总的 A 出现的次数。可以看出,基于区分矩阵的属性约简策略很耗费时间。设这23个对象中取 H 的 x_1 个,取 M 的 x_2 个,取 LO 的 x_3 个,那么上述问题求取 A 出现的次数的问题就是求取 $x_1\times x_2+x_1\times x_3+x_2\times x_3$ 的问题。所以,如这23个对象的 $u_1,u_2,u_3,\cdots,u_{23}$ 的取值 H, H, H, H, M, M, H, LO, LO, M, LO, M, M, H, LO, H, H, H, M, M, M, M, M,按照本文的解法,只要计算出这23个对象中取值 H 的为9个,取值 M 的为10个,取值 LO 的为4个,所以 A 出现的次数为 $9\times10+9\times4+10\times4=140$。

研究基于区分矩阵的约简策略,可以发现,在求取第一个重要属性的过程中,第一个重要属性就是在区分矩阵中出现的次数最大的属性。设信息表 S 有 n 个实例 m 个属性,属性值 $A=(A_1,A_2,\cdots,A_p)$,对于属性 $j(j=1,2,\cdots,m)$,对应属性值 v 的个数向量 $x_i=(x_{j1},x_{j2},\cdots,x_{jp})$。那么寻求第一个重要属性的过程,使 $x_{j1}\times x_{j2}+x_{j1}\times x_{j3}+\cdots+x_{j(p-1)}\times x_{jp}$ 达

*)国家自然科学基金资助项目(批准号:60573056),浙江省自然科学基金资助项目(批准号:Y105090)。**谢晓飞** 讲师,硕士,主要研究方向为模式识别、中文信息处理及软件工程;**邵 斌** 副教授,硕士,主要研究方向为人工智能和数据融合;**张建宏** 讲师,硕士,主要研究方向为人工智能和数据融合。

到最大的属性就是最重要属性，记 $x_{j1}\times x_{j2}+x_{j1}\times x_{j3}+\cdots+x_{j(p-1)}\times x_{jp}$ 为属性函数 ArF。所以，计算各属性的 ArF，就可以定出第一个重要属性，此时重要属性的 ArF 最大。

在第一个重要属性成功地简单求出后，研究基于区分矩阵的约简策略，可以发现区分矩阵的实质是建立在两两相异的对象对上属性集合的绝对差异矩阵。所以下一个重要属性应是剩余的属性中的 ArF 减去重要属性中已包含的两两相异对后的最大属性。设 ArM 为属性矩阵，其元素 ArM(a_j) 为属性 a_j 的所有两两相异对的集合。

2 近似约简算法

经过属性的约简后，属性的个数是减少了，但在大部分的信息系统中，属性的个数仍然很多，且有些属性之间的差异是由于数据的噪音所引起的，故对于这些属性可以进一步约简，而又能基本保持原信息系统的风格面貌。

在 ARS 算法中，建立了属性的两两相异对的集合，利用该集合可找出相对重要的属性，要找出了相对于完全约简的属性集较小的一个属性集合后，剩余的两两相异对就相对很少了，而这些属性的略去在一定程度上不会影响到原系统的信息。

设信息系统的 n 个对象总体 U，它可取 r 个可能的值 $\{u_1,u_2,\cdots,u_r\}$，每个取值的样本数为 $\{p_1,p_2,\cdots,p_r\}$，在约简了几个属性后，每个取值的样本数为 $\{p_1',p_2',\cdots,p_r'\}$，现在要检验：$H_0:p_i=p_i'\quad i=1,2,\cdots,r$。

由数理统计的大数定理可知，当 H_0 成立时，p_i 与 p_i' 应相差不大，反之，如两者相差很大，那么自然就有理由认为 H_0 不成立时。据此 Pearson 曾构造出反映这种差异的统计量：

$$\chi^2\sim\sum_{i=0}^{i}\frac{(p_i'-p_i)^2}{p_i}$$

并证明了当 H_0 成立且 n 充分大时，不论总体服从什么分布，检验统计量 χ^2 总是近似地服从自由度为 $r-1$ 的 χ^2 分布。

定义 4(近似约简)　对于 $x(x\in U)$，一个集合 $R\subseteq A$ 是信息系统 S 关于 x 的一个近似约简，若 $S_R(x)$ 的分布与 $S_A(x)$ 的分布对于 χ^2 统计量 $\sum_{i=0}^{i}\frac{(p_i'-p_i)^2}{p_i}$ 在显著水平 α 下服从相同分布，且对于 $\forall B\subset R$，$S_B(x)$ 的分布与 $S_A(x)$ 的分布对于 χ^2 统计量在显著水平 α 下不服从相同的分布。

令 ArM 是信息表 S 中属性的约简矩阵，ArM 的元素 $RM(a_k)=\{u_i,u_j\}$，它是信息表中属性 a_j 的所有两两相异对象的集合。建立属性判断 ArF，ArF 的元素 $RM(a_k)$，它对应于属性 a_k 的所有两两相异对象的集合个数。

故属性近似约简算法描述如下：

令 $R=\varnothing$，计算 ArM，对信息表形成初始分类，r 为初始分类数，计算每个类别的样本数 $p=\{p_1,p_2,\cdots,p_r\}$。

循环：

(1)计算 ArF，有 ArF(a_k)=Max(a_j)

(2) $R<-R\cup\{a_k\}$

(3)令 ArM<--ArM(a)-ArM(a_k)

(4)根据 ArM 将对象分类

①扫描 ArM，对每个$\{u_i,u_j\}$，将对象 u_j 合并到 u_i 中。

②重新计算每个类别的样本数 $p'=\{p_1',p_2',\cdots,p_r'\}$。

(5)计算 $\chi^2\sim\sum_{i=0}^{i}\frac{(p_i'-p_i)^2}{p_i}$ 统计量

直到 $\chi^2<\chi_\alpha^2(r-1)$

此时集合 R 是信息表在显著水平 α 下的一个属性近似约简。

3 实例分析

对于一个如表 1 所示的信息系统约简过程如下：

表 1　信息系统

U/A	a1	a2	a3	a4	a5	a6	a7	a8	a9
1	C	6	Y	E	M	H	H	A	M
2	C	6	N	E	M	M	H	MA	M
3	C	6	N	E	M	H	H	MA	M
4	C	4	Y	E	M	H	H	MA	L
5	C	6	N	E	M	M	M	MA	M
6	C	6	N	B	M	M	M	A	HE
7	C	6	N	E	M	M	H	MA	HE
8	S	4	N	B	SM	H	LO	MA	L
9	C	4	N	B	SM	H	LO	MA	M
10	C	4	N	B	SM	H	M	A	M
11	S	4	N	E	SM	H	LO	MA	L
12	S	4	N	E	M	M	M	MA	M
13	C	4	N	B	M	M	M	MA	M
14	S	4	Y	E	SM	H	H	MA	M
15	S	4	N	B	SM	M	LO	MA	M
16	C	4	Y	E	M	M	H	MA	M
17	C	6	N	E	M	M	H	A	M
18	C	4	N	E	M	M	H	A	M
19	S	4	N	E	SM	H	M	MA	M
20	C	4	N	E	SM	H	M	MA	M
21	C	4	N	B	SM	H	M	MA	M
22	C	6	N	B	M	M	M	A	HE
23	C	4	N	B	SM	H	M	A	M

如对于如表 1 的信息系统根据 ARS 算法进行约简，$R=\{a_7,a_6,a_8,a_1,a_2,a_4,a_9\}$ 为原信息系统的属性约简。

如对于如表 1 的信息系统进行近似约简，显著水平 α 取 0.05，初始分类为 $\{\{u_1\},\{u_2\},\{u_3\},\{u_4\},\{u_5\},\{u_6,u_{22}\},\{u_7\},\{u_8\},\{u_9\},\{u_{10},u_{23}\},\{u_{11}\},\{u_{12}\},\{u_{13}\},\{u_{14}\},\{u_{15}\},\{u_{16}\},\{u_{17}\},\{u_{18}\},\{u_{19}\},\{u_{20}\},\{u_{21}\}\}$，各类的样本数为{1,1,1,1,1,2,1,1,1,2,1,1,1,1,1,1,1,1,1,1,1}，$r=21$，查 χ^2 分布表，$\chi_\alpha^2(r-1)=31.40$，根据 ARS 算法，当计算到 $R=\{a_7,a_6,a_8\}$ 时，ArM(属性约简矩阵)中 $RM(a_1)$ 为{{12,13},{19,20},{19,21},{3,14},{4,14},{5,12},{8,9},{9,11}}，$RM(a_2)$ 为{{17,18},{2,16},{3,14},{3,4},{5,12},{5,13},{7,16}}，$RM(a_3)$ 为{{2,16},{3,14},{3,4},{7,16}}，$RM(a_4)$ 为{{12,13},{19,21},{20,21},{5,13},{8,11},{9,11}}，$RM(a_5)$ 为{{3,14},{4,14}}，$RM(a_9)$ 为{{2,7},{3,4},{4,14},{7,16},{8,9},{9,11}}。根据此 ArM 表可知原信息表中只有 $\{a_7,a_6,a_8\}$ 三个属性时，对象{2,7,16}、对象{3,4,14}、对象{5,12,13}、对象{8,9,11}、对象{17,18}、对象{19,20,21}将被分别分在一类中，故此时的分类为：$\{\{u_1\},\{u_2,u_7,u_{16}\},\{u_3,u_4,u_{14}\},\{\},\{u_5,u_{12},u_{13}\},\{u_6,u_{22}\},\{\},\{u_8,u_9,u_{11}\},\{\},\{u_{10},u_{23}\},\{\},\{\},\{\},\{\},\{u_{15}\},\{\},\{u_{17},u_{18}\},\{\},\{u_{19},u_{20},u_{21}\},\{\},\{\}\}$。各类的样本数：{1,3,3,0,3,2,0,3,0,2,0,0,0,0,1,0,2,0,3,0,0}，$\chi^2=\frac{(1-1)^2}{1}+\frac{(3-1)^2}{1}+\frac{(3-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(3-1)^2}{1}+\frac{(2-2)^2}{2}+\frac{(0-1)^2}{1}+\frac{(3-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(2-2)^2}{2}+\frac{(0-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(1-1)^2}{1}+(\frac{(0-1)^2}{1}+\frac{(2-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(3-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(0-1)^2}{1}=32>\chi_\alpha^2(r-1)=31.40$，故认为此分类与原分类不服从相同的分布。

这时 ArF(a_1)最大，故 $R=\{a_7,a_6,a_8,a_1\}$，这时 ArM 中，

$RM(a_2)$为{{17,18},{2,16},{3,4},{5,13},{7,16}},$RM(a_3)$为{{2,16},{3,14},{7,16}},$RM(a_4)$为{{20,21},{5,13},{8,11}},$RM(a_5)$为{},$RM(a_9)$为{{2,7},{3,4},{7,16}}。根据此ArM表可知原信息表中只有$\{a_7,a_6,a_8,a_1\}$四个属性时,对象{2,7,16}、对象{3,4}、对象{5,13}、对象{8,11}、对象{17,18}、对象{20,21}将被分别分在一类中,故此时的分类为:$\{\{u_1\},\{u_2,u_7,u_{16}\},\{u_3,u_4\},\{\},\{u_5,u_{13}\},\{u_6,u_{22}\},\{\},\{u_8,u_{11}\},\{u_9\},\{u_{10},u_{23}\},\{\},\{u_{12}\},\{\},\{u_{14}\},\{u_{15}\},\{\},\{u_{17},u_{18}\},\{\},\{u_{19}\},\{u_{20},u_{21}\},\{\}\}$。各类的样本数:{1,3,2,0,2,2,0,2,1,2,0,1,0,1,1,0,2,0,1,2,0},$\chi^2=\frac{(1-1)^2}{1}+\frac{(3-1)^2}{1}+\frac{(2-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(2-1)^2}{1}+\frac{(2-2)^2}{2}+\frac{(0-1)^2}{1}+\frac{(2-1)^2}{1}+\frac{(1-1)^2}{1}+\frac{(2-2)^2}{2}+\frac{(0-1)^2}{1}+\frac{(1-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(1-1)^2}{1}+\frac{(1-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(2-1)^2}{1}+\frac{(0-1)^2}{1}+\frac{(1-1)^2}{1}+\frac{(2-1)^2}{1}+\frac{(0-1)^2}{1}=16<\chi_\alpha^2(r-1)=31.40$,故认为此分类与原分类服从相同的分布。

所以这时$R=\{a_7,a_6,a_8,a_1\}$即为原信息系统的一个近似约简。

结束语 本文出的近似约简算法的时间复杂度仍为$O(n)$,但其约简后的属性大大减少。本例中原有9个属性,经过属性约简后剩7个属性,但经过本算法近似约简后的属性仅为4个,在显著水平$\alpha=0.05$下与原信息系统服从相同的分布。这样的约简使大信息系统在基本保留原风格的情况下尽可能少地保留属性,即尽可能多地约简属性,为后期的系统处理节约了大量的处理时间。

参 考 文 献

1 Pawlak Z. Rough Sets. International Journal of Information and Computer Sciences, 1982, 11:341~356
2 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Boston, London, Dordrecht, 1991
3 Pawlak Z, et al. Rough Sets. Communications of the ACM, 1995, 38(11):89~95
4 Blum A, Langley P. Selection of Relevant Feature and Examples in Machine Learning. Artificial Intelligence, 1997,72:245~271
5 Almuallim H, Dietterich T. Learning Boolean Concepts in the Presence of Many Irrelevant Features. Artificial Intelligence, 1994, 69(1-2):279~305
6 Kira K, Rendell L. The Feature Selection Problem: Traditional Methods and A New Algorithm. In:Proceedings of the Tenth National Conference on Artificial Intelligence, Menlo Park, AAAI Press/The MIT Press, 1992. 129~134
7 Modrzejewski M. Feature Selection Using Rough Sets Theory. In:Proceedings of the European Conference on Machine Learning, Vienna, 1993. 213~226
8 Skowron A. The Discernibility Matrices and Functions in Information Systems, Intelligent Decision Support-Handbook of Applications and Advances of Rough Sets Theory, Kluwer Academic Publishers, Dordrecht, Boston, London, 1992. 331~363
9 秦中广.基于粗糙集的交叉研究及其在中医诊断的应用:[博士论文].华南理工大学,2002.15~29
10 张文修,等.粗糙集理论与方法.北京:科学出版社,2001
11 赵选民,等,编.数理统计.北京:科学出版社,2002

(上接第98页)

拽方式对BPEL流程建模。在模型服务器的BPEL引擎执行流程前,需要生成一个流程描述符,它特定于BPEL服务器,必须重写它才能在不同BPEL引擎上运行BPEL流程。本模型服务器的流程描述符是包括BPEL流程ID,BPEL源文件名、合作伙伴链接的WSDL文件等,是一个XML文件。

(5)集成调度服务器:调度服务,可视化服务配置和运行状态,判断各种空间信息资源的稳定度、是否在线,为LBS应用服务器的服务选择时提供服务的服务质量(QoS)信息。在模型服务器执行流程实例过程中,集成调度服务器能将流程中各服务的执行状态(如正在运行、已完成)信息传给模型服务器。

在此原型系统中,面对不同LBS应用需求,在LBS应用服务器只需处理应用逻辑,功能模块通过对服务平台的服务调用实现,各服务节点提供的基本服务可以满足应用需求的,就直接调用该子服务,对复杂的服务的调用,通过模型服务器对流程模型的执行获得。这样一方面充分利用了共享的服务资源,另外一方面确保了应用系统开发的灵活性和快速性。

结束语 在LBS应用中对空间数据有着日益增加的需求,绝大部分应用系统不具备完善和快速更新空间数据的能力,Web服务接口所定义的消息与平台和语言无关,均可与Web服务程序通信,进行数据交换,当LBS应用中第三方数据以Web服务的方式提供,就屏蔽了底层异构数据管理的复杂性,可实现数据的快速共享。同时,在LBS中对移动终端位置的定位技术也较多,按照国际相关接口协议将其封装为Web服务也减少了系统开发的难度。本文的实践证明,在LBS应用系统中将常见的数据访问、分析处理、计算通过服务平台集成起来,并采用服务组合技术灵活的根据业务逻辑构造新的服务,注册后为各类的LBS应用提供可访问的服务接口,具有很好的开放性。服务平台的不同服务节点会提供相似服务,无论应用系统对子服务的直接调用,或是复杂服务对子服务的组合都需要对其动态选择,对服务语义匹配有严格的要求在本文实践中,假设服务选择中待选的相似服务是严格匹配的,按照服务的QoS对服务进行了动态选择。在分布式的环境下,各子服务的状态是确保系统正确执行的关键,当一个服务出现问题,备用服务的平滑的动态切换也是进一步需要解决的问题。

参 考 文 献

1 岳昆,王晓玲,周傲英. Web服务核心支撑技术:研究综述. 软件学报,2004,15(3)
2 Tosic V, Mennie D, Pagurek B. On Dynamic Service Composition and Its Applicability to E-Business Software Systems. WOOBS'01 (Workshop on Object-Oriented Business Solutions), Budapest ,2001
3 OGC. OpenGIS Location Services (OpenLS): Core Services. http://portal.opengeospatial.org
4 OMA. http://www.openmobilealliance.com/tech/affiliates/lif/lifindex.html
5 唐宇,何凯涛,陈荦,等. 空间信息栅格体系与服务聚合.国防科学技术大学学报,2005,27(2)
6 王桂玲,李玉顺,姜进磊,等. 一种服务网格动态信息聚合模型及其应用. 计算机学报,28(4)
7 Velasco Juan R,Castillo Sergio F. Mobile agents for web service composition. Lecture Notes in Computer Science,2003
8 Lee J. Matching algorithms for composing business process solutions with web services. Lecture Notes in Computer Science,2003
9 Kim J-C, Heo T-W, et al. Ubiquitous Location Based Service. In: Proceedings of the 8th International. IEEE Conference on Intelligent Transportation Systems,Vienna,2005