基于粒度计算的特征选择方法*)

孙丽君　苗夺谦

(同济大学计算机科学与技术系　上海 201804)

摘　要　从粒度计算的划分模型出发,重新定义了相容决策表的约简,并给出了一种新的基于粒度计算的属性约简算法。该算法以信息熵作为启发信息,通过逐渐增加属性构成条件属性集相对于决策属性的约简,再通过删除约简中的所有不必要属性,得到最小约简。该算法有效地降低了计算属性约简的时间复杂度,可以用于较大规模数据集的特征选择。在5个公开的基因表达数据集上的实验证明了该算法能找到高区分能力的特征子集。

关键词　粒度计算,粗糙集,约简,特征选择

Method for Feature Selection Based on Granular Computing

SUN Li-Jun　MIAO Duo-Qian

(Department of Computer Science and Technology, Tongji University, Shanghai 201804)

Abstract　The reduction of consistent decision table is redefined from the point of view of partition module of granular computing, and a novel algorithm to find an optimal reduction with low time complexity is proposed. Reduction is constructed by adding attributes using information entropy as the heuristic information, superfluous attributes are deleted from the reduction to get a minimal reduction. The experiment results on five public gene expression profiles demonstrate the usefulness of this new method for feature selection on large scale data sets.

Keywords　Granular computing, Rough sets, Reduction, Feature selection

1　引言

特征选择是机器学习领域的重要课题,其任务是从全部的属性集合中挑选出对学习任务重要的属性,删除冗余的和不必要的属性。近年来,特征选择方面的研究引起机器学习领域学者的高度重视,主要原因有以下两个方面:1)大多数学习算法所需训练样本的数目随不相关特征的增多而急剧增加,因此选择好的特征子集不仅可以减小计算复杂度,提高分类准确度,而且有助于寻找更精简、更易理解的算法模型。2)大规模数据处理问题的不断出现使得数据挖掘的发展对大规模数据处理的研究提出了迫切的要求,如信息检索、遗传基因分析等。

Pawlak提出的粗糙集理论提供了寻找所有可能的特征子集的数学工具[1]。粗糙集理论中,数据存放在决策表 $T=(U,C\cup D)$ 中,且 $C\cap D=\phi$,U 是论域,C 和 D 分别对应着条件属性集和决策属性集。多数的基于粗糙集的特征子集选择算法都是面向属性约简的。也就是说,寻找条件属性集的最小约简 R,它与 C 有着相同的分类能力,因此我们就可以用 R 来代替 C 挖掘分类规则。决策表可能有不止一个的约简,其中的任何一个都可以用来替代原始的属性集合 C。现有的粗糙集约简算法往往需要根据所有条件属性和决策属性的值去构造等价类,算法的时间复杂性都较高,一般为 $O(n^2)$,n 为属性的个数。粗糙集的约简算法的低效性限制了粗糙集理论的广泛应用,特别是在一些大规模的数据集上的应用。

粒度计算是描述问题空间和解决问题的有效手段,它使我们能在不同的粒度上洞察现实世界,获取有用的知识[2~7]。其核心思想是构造粒子,并表示它们的关系。也就是说,初始的问题空间被划分成一些基本粒子,这些基本粒子在不同的粒度层次上再组合或分解成新的粒子。这样反复执行,直到新的粒子能更有效地解决问题。粗糙集理论是粒度计算领域的主要研究方向之一,很多粗糙集理论的应用可以被看作粒度计算的具体模型。关于决策表,不同粒度观点内之间的映射事实上由不同的属性子集来定义,粒度之间的转换通过增加删除属性来实现,因而寻找属性约简的过程可以模型化为寻找由属性的所有子层定义的层级划分。

在本文中,我们从粒度计算的划分模型出发,提出了一种新的基于粒度计算划分模型的决策表属性约简算法,该算法有效地降低了计算属性约简的时间复杂性,可以用于较大规模的数据集的特征选择。

2　粒度计算的划分模型

决策表中,论域 U 的一个划分 $\pi=\{X_i\mid 1\leqslant i\leqslant m\}$ 对论域提供了简单的粒度观点,划分中的每个块 X_i 都是一个粒子,且满足:

- 每个 X_i 非空
- 对所有的 $i\neq j$,$X_i\cap X_j=\phi$
- $\cup\{X_i\mid 1\leqslant i\leqslant m\}=U$

如果划分 π_1 的所有块都包含在划分 π_2 的块中,称划分 π_1 是 π_2 的细化,记作 $\pi_1\leqslant\pi_2$。细化关系是一个偏序,满足自反性、对称性和传递性,它定义了一个划分格 $\Pi(U)$。$\Pi(U)$ 包含了论域的所有可能的粒度划分,在划分上的细化偏序关系则提供了一个自然的分层递阶结构。空集提供了最粗的划分,全部的属性定义的划分是最细的划分,粒度之间的转换通过增加删除属性来实现。Y. Y. Yao在文[5]中,从粒度计算

*)国家自然科学基金资助项目(60475019)。孙丽君　讲师,博士生。

的划分模型出发，重新定义了相容决策表的核属性、不必要属性和约简。

定义 1 相容决策表 $T=(U,C\cup D)$ 中对于 $\forall c\in C$，如果不满足 $\pi_{C-\{c\}}\leqslant\pi_{C-\{c\}\cup D}$，则称 c 是 C 中相对于 D 的核属性，否则 c 是 C 中相对于 D 的不必要属性。核属性是数据集中不可缺少的属性，它包含了用于分类的必要信息；不必要属性是一些冗余的属性，完全可以从数据集中删除而不会影响分类的结果。

定义 2 相容决策表 $T=(U,C\cup D)$ 中，C 的子集 R 是 C 的约简，当且仅当 $\pi_R\leqslant\pi_D$，并且 $\forall Q\subset R$，π_Q 不满足 $\pi_Q\leqslant\pi_D$。

决策表 $T=(U,C\cup D)$ 中，假设 $R\subseteq C$，R 上的不可区分关系 $IND(R)=\{(x,y)\in U\times U\mid\forall a\in R,a(x)=a(y)\}$ 也是 U 上的一个等价关系，它构成了对 U 的一个划分，记为 $\pi_{IND(R)}$。其中的每一个等价类 $[x]_{IND(R)}$ 就可以看作是一个可定义的粒子。如果 $\pi_{IND(R)}$ 是比由决策属性集上的不可区分关系 $IND(D)$ 导出的划分 $\pi_{IND(D)}$ 更细的划分，即 $\pi_{IND(R)}\leqslant\pi_{IND(D)}$，对相容决策表而言，可以很容易地证明 $|\pi_{IND(R)}|=|\pi_{IND(R\cup D)}|$。根据定义 1,2，我们可以得到相容决策表核属性、不必要属性和约简的新定义：

定义 3 相容决策表 $T=(U,C\cup D)$ 中对任意属性 $c\in C$，如果满足 $|\pi_{IND(C-\{c\})}|\neq|\pi_{IND(C-\{c\}\cup D)}|$，则 c 是 C 中相对于 D 的核属性；如果满足 $|\pi_{IND(C-\{c\})}|=|\pi_{IND(C-\{c\}\cup D)}|$，则 c 是 C 中相对于 D 的不必要属性。

定义 4 相容决策表 $T=(U,C\cup D)$ 中，如果 $R\subseteq C$ 满足 $|\pi_{IND(R)}|=|\pi_{IND(R\cup D)}|$ 并且 $\forall Q\subset R$ 满足 $|\pi_{IND(Q)}|\neq|\pi_{IND(Q\cup D)}|$，$R$ 就是 C 相对于 D 的约简。

3 基于粒度计算的属性约简算法

基于上述定义 3,4，我们给出了求解决策表属性约简的一种启发式算法。该算法通过逐渐增加属性构成条件属性集 C 相对于决策属性 D 的约简，再通过删除约简中的所有不必要属性，得到最小约简。算法中信息熵作为启发信息，步骤如下：

1) $RED(C)\leftarrow\phi$，计算所有 $q\in C-RED(C)$ 的熵 $E(q)$；

2) 判断 $|\pi_{RED(C)}|=|\pi_{RED(C)\cup D}|$?，若成立转 5)，否则转 3)；

3) 取 q_1 满足 $E(q_1)=\min\limits_{q\in C-RED(C)}E(q)$；

4) $RED(C)=RED(C)\cup\{q_1\}$，把属性 q_1 加入到约简中。转 2)；

5) 对每一个 $q\in RED(C)$，如果 $|\pi_{RED(C)-\{q\}}|=|\pi_{RED(C)-\{q\}\cup D}|$，$RED(C)=RED(C)-\{q\}$；

6) 输出最小约简 $RED(C)$。

算法中，任一属性 $q\in C$ 的熵 $E(q)$ 按下式计算：

$$E(q)=\sum_{Y\in\pi_{IND(q)}}\frac{|Y|}{|U|}I(q/Y)=-\frac{1}{|U|}\sum_{X\in\pi_{IND(D)}}\sum_{Y\in\pi_{IND(q)}}|X\cap Y|\log_2\frac{|X\cap Y|}{|Y|}$$

计算的时间复杂性为 $O(|U|\log|U|)$，因而步骤 1) 的计算复杂性为 $O(|C||U|\log|U|)$。若我们先对数据集按属性子集 B 进行排序，只要扫描数据集一遍就可以求得对应的等价类。根据属性子集 B 对 $|U|$ 个对象进行排序的算法复杂性为 $O(|B||U|\log|U|)$，则步骤 2) 和 5) 的计算复杂度为 $O(|M||U|\log|U|+|U|)$。

其中 $|M|$ 为约简中的属性数。整个算法的复杂性为

$O(|C||U|\log|U|+|M||U|\log|U|+|U|)$

因为 $|M|\ll|C|$，所以算法的复杂度是 $O(|C||U|\log|U|)$。

4 实验

基因表达数据集可以归为决策表，其中的每一个记录对应着的一个样本，条件属性集包含每一个基因，决策属性对应着癌症的子类。基因表达数据集的特点是属性数目较多，有几千个甚至上万个。在众多的属性中，可能只有少数的属性与目标症状有较强的相关性，因而对基因表达数据进行特征选择是十分必要的。我们选用 5 个公开的基因表达数据集，在 Pentium 1.8GHz PC(1.0GB RAM, WINXP) 上进行了约简，以验证我们的算法的有效性。实验数据集见表 1。

表 1 实验所用数据集

数据集	属性数	样本数	决策类
ALL_AML LEUKEMIA	7069	72	2
COLON CANCER	2000	62	2
LUNG CANCER	12533	181	2
CENTRAL NERVOUS SYSTEM	7129	60	2
MLL_LEUKEMIA	12582	72	3

因为粗糙集方法只能处理离散的数据，在约简之前我们先要对数据集进行离散化。实验中我们使用了文[8]中的简单离散化算法对基因表达数据进行离散化，对离散化后的基因表达数据集应用上述约简算法进行特征选择，找出对疾病分类有较强区分能力的基因特征子集。

为验证选择出的基因子集对疾病分类的有效性，约简对应的规则集用作分类器，来预测未知样本的类别。10-折交叉检验用于检验分类器的有效性。实验结果见表 2。

对这 5 个基因表达数据集，我们挑选出的特征子集均包含不到 10 个的属性，但是均达到了高于 90% 的预测准确率。这说明我们的算法能选择出高区分能力的特征子集，从而降低分类算法的计算代价。从实验结果中我们可以看出，具有 12533 个属性的 LUNG CANCER 数据集和具有 12582 个属性的 MLL _LEUKEMIA 数据集的分类正确率明显高于其他几个数据集。这说明随着数据集中属性数的增加，我们的算法选择出的特征子集的区分能力越强。

表 2 实验结果

数据集	属性数	样本数	约简后属性数	运行时间	预测正确率
ALL_AML LEUKEMIA	7069	72	5	8s	92.9%
COLON CANCER	2000	62	7	2s	91.7%
LUNG CANCER	12533	181	7	62s	97.8%
CENTRAL NERVOUS SYSTEM	7129	60	7	6s	91.7%
MLL_LEUKEMIA	12582	72	6	15s	97.1%

结论 粗糙集理论提供了寻找所有可能的特征子集的数学工具，但是现有的粗糙集约简算法的低效性限制了粗糙集理论的广泛应用，特别是在一些大规模的数据集上的应用。本文中，我们从粒度计算的划分模型出发，重新定义了相容决策表的约简，并给出了一种基于粒度计算的属性约简算法。该算法有效地降低了计算属性约简的时间复杂度，可以用于较大规模数据集的特征选择。在 5 个公开的基因表达数据集上应用该算法，我们都得到了不足 10 个属性的特征子集，特征子集导出的分类规则对原数据集进行分类，均取得了较高

（下转第 39 页）

(3)测试数据包中某数据记录的离散型属性向量关联规则,不在训练数据包所得的离散型属性向量关联规则中。

对此,本文实验研究中将出现概率较高的离散型属性向量关联规则,附加上部分连续型属性规则。例如,某离散型属性向量关联规则只在正常网络连接数据记录中出现,具有独有性,而实际情况中不断出现的新攻击类型,总是会违反正常网络连接数据记录的某些特性。基于此,若某网络连接数据记录的离散型属性向量关联规则与正常离散型属性向量关联规则相同,但违反某正常连续型属性向量规则,且不在已知攻击类型的相应连续型属性向量规则范围内,则该数据记录为新的攻击类型,或属于某已知攻击类型。

4.3 特征匹配与聚类算法的协同检测

表7 特征匹配与聚类算法协同检测时测试数据包的检测结果

类型	具体类型	测试包中数据量	正确检测数据量	检测率
DOS	back	1098	1098	100%
	land	9	9	100%
	neptune	58001	52280	90.1%
	pod	87	87	100%
	smurf	164091	163772	99.8%
	teardrop	12	10	83.3%
Probing	ipsweep	306	253	82.7%
	nmap	84	80	95.2%
	portsweep	354	286	80.8%
	satan	1633	1312	80.3%
R2L	ftp_write	3	2	66.7%
	guess_passwd	4367	3673	84.1%
	imap	1	1	100%
	multihop	18	13	72.2%
	phf	2	1	50%
	spy	无		
	warezclient			
	warezmaster	1602	1063	66.4%
U2R	buffer_overflow	22	12	54.5%
	loadmodule	2	2	100%
	perl	2	1	50%
	rootkit	13	10	76.9%
normal		60593	60223	99.4%

将连续型属性规则的特征匹配,与基于分箱统计的FCM算法结合,其检测过程分以下两种情况:

(1)当新数据记录的离散型属性向量组成的离散型关联规则在离散型关联规则集中为独有时,计算该新数据记录到此离散型关联规则对应聚类的距离,使用式(3)判断新数据记录是否属于该聚类。

①如果属于该聚类,则与该聚类的分箱比较,判断是否要更新该聚类。

②如果不属于该聚类,则根据该新数据记录建立一新的聚类。

(2)当新数据记录的离散型属性向量组成的离散型关联规则在离散型关联规则集中为非独有时,与有该规则的聚类的连续型规则比较,判断在哪个聚类的连续型规则范围内。

①当只在一个聚类的连续型规则内时,该新数据记录属于该聚类,并计算到该聚类的距离,判断是否要更新该聚类。

②当在几个聚类的连续型关规则内时,计算到这些聚类的距离,使用式(3)判断新数据记录属于哪个聚类,并与该聚类的分箱比较,判断是否要更新该聚类。

③当不在任何已知聚类的连续型关规则内时,该新数据记录为新的类型,为此数据记录建立一新的聚类。

对测试数据包进行实验,其结果如表7所示,总检测率达97.2%。实验证明检测引擎上使用特征匹配与聚类算法的协同分析检测,不但速度快,正确性较高,且便于自动更新检测模块,发现新的攻击类型。

结束语 在网络入侵检测数据实验中,基于分箱统计的FCM聚类算法对部分聚类的检测率较为理想,但整体检测效果并不好。其原因可能在于:离散型属性向量关联规则在多个聚类中存在时,往往将相关的数据记录判为关联规则对应的离散型属性向量概率高的聚类;当某聚类的概率值较高时,与该关联规则相关的聚类的部分数据记录可以正确检测出来,其余部分检测判断错误;测试数据包中某数据记录的离散型属性向量关联规则,不在训练数据包所得的离散型属性向量关联规则中。

将特征匹配与基于分箱的FCM算法相结合,协同分析网络连接数据记录,这时检测率有明显提高,实时性好,能较好地发现新的攻击类型,便于检测知识库的更新。

参考文献

1 高新波著.模糊聚类分析及其应用.西安:西安电子科技大学出版社,2004.1

2 Ramaswamy S,Rastogi R,Shim K. Efficient algorithms for mining outliers from large data sets. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 2000. 427~438

3 Portnoy L,Eskin E,Stolfo S J. Intrusion detection withunlabeled data using clustering. In: Proceedings of theACM Workshop on Data Mining Applied to Security,Philadelphia, PA, 2001

4 Sequeira K, Zaki M. ADMIT: Anomaly-based data mining for intrusions. In: Proceedings of the 8th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, 2002. 386~395

5 http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.htm

(上接第15页)

的分类正确率,这表明了该算法可以在较大规模的数据集上进行有效的特征选择。

参考文献

1 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic publishers,1991

2 Liu Q. Granules and Application of Granular Computing in Logical Reasoning. Chinese of Computer Research and Development, 2004, 41(4):546~551

3 Yao Y Y. A Partition Model of Granular Computing. T Rough Sets,2004.232~253

4 李道国,苗夺谦,张红云.粒度计算的理论、模型与方法.复旦学报, 2004, 43(5): 837~841

5 Yao Y Y,Yao J T. Granular computing as a basis for consistent classification problems. In: Proceedings of PAKDD'02 Workshop on Toward the Foundation of Data Mining, Taipei,2002. 101~106

6 邓蔚,王国胤,吴渝.粒计算综述.计算机科学,2004, 31(z2):178~181

7 Yao Y Y. Three Perspectives of Granular Computing. 南昌工程学院学报,2006,25(2): 16~20

8 Ding C, Peng H C. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. In: Proceedings of the Computational Systems Bioinformatics , California ,2003