最大泛化规则生成

Generation of Maximally Generalized Rules

徐如燕[1] 鲁汉榕[2] 郭齐胜[1]

(装甲兵工程学院 北京100072)[1] (空军雷达学院 武汉430010)[2]

Abstract In this paper, the generation of maximally generalized rules in the course of classification knowledge discovery based on rough sets theory is discussed. Firstly, an algorithm is introduced. Secondly, we propose that the information-based J-measure is used as another measure of attribute significance value. This measure is used for heuristically selecting the conditions to be removed in the process of extracting a set of maximally generalized rules. Finally, we present an example to illustrate the process of the algorithm.

Keywords Classification knowledge discovery, Maximally generalized rules, Information theory, J-measure, Significance

1 引言

在基于粗糙集理论的分类知识发现中,用产生式规则来表示最终的发现结果,称为分类规则。一般形式为:

$r:$ if $c_1 \& c_2 \& \cdots \& c_n$ then D

其中,c_i 是条件属性,D 是决策属性。在将初始的信息表进行属性泛化处理并计算约简之后,接下来的工作就是由约简生成分类规则[1]。一个约简的信息系统可被看作是一组特定的规则,每个元组可以直接写成一条逻辑规则。

我们的目标是在学习过程中产生最大泛化规则,这是通过尽可能多地删除条件属性值而不降低规则的分类精确度而得到的,计算最大泛化规则在数据挖掘的应用中特别重要,因为它们表示了数据中存在的最泛化模式。

本文首先介绍最大泛化规则生成的算法。在这个算法中,每次选择相对重要性较小的条件进行删除,同时通过不断检验规则的一致性来保证规则的分类精确度。因此,需要为每个条件属性赋予一个反映其与决策属性相关性的"有效度"来确定删除的顺序。一般采用概率的方法[1]来计算有效度,但使用这种方法计算出的有效度可能出现负值,而且可能因受事件出现频率的影响而不能反映真实的信息量。本文提出了以基于信息的 J-measure 作为有效度量的方案,阐明了它优于以往所使用的有效度量之处,并用实例说明了最大泛化规则生成算法的执行过程。

2 最大泛化规则的生成

分类规则发现是一种较高概念层次上的知识发现,它所发现的最终结果是一些最大泛化规则。生成最大泛化规则的过程,是在信息表约简的基础上,对每条由元组直接写成的规则,尝试删除条件属性,在不降低规则的分类精确度的前提下,尽可能地缩短规则长度。

算法1 计算一组最大泛化规则[1]。

输入:一个具体决策规则的非空集合 RULE

输出:一个最大泛化规则的非空集合 MRULE

```
MRULE←0; N←|RULE|; /* N 是 RULE 中的规则数 */
for i=0 to N-1 do
  r←r_i;
  M←|r|; /* M 是规则 r 的条件属性数 */
  对规则 r 的每个条件,计算其有效度 SIG;
  将规则的条件集合按 SIG 排序;
  for j=0 to M-1 do
    删除规则 r 的第 j 个条件属性 c_j;
    if r 与任意的规则 r_n∈RULE 不一致 then
      恢复条件 c_j
  endif;
    endfor;
    删除逻辑包含于规则 r 的所有规则 r'∈MRULE;
    if 规则 r 不逻辑包含于任一规则 r'∈MRULE then
      MRULE←r∪MRULE
    endif;
endfor.
```

算法1考虑具体决策规则集中的每个规则,删除条件,直至得到一组最大泛化规则。算法中每删除一条规则的一个属性,就要检验所生成的新规则是否与原规则集一致,若不一致,则立即恢复这个属性,从而保证了规则的分类精确度不降低。

以不同的顺序处理属性,会生成不同的最大泛化规则。因此,一条最大泛化规则从简洁度或覆盖度的角

度看不一定是最佳的。假设一条规则有 m 个条件，我们可以在数据库中考察2^m-1个可能的条件子集，选出最佳的规则。但是，当规则中的条件属性数很大时，生成最优的最大泛化规则通常是不可行的。一种启发式的解决方法是，在删除条件的操作开始之前，给规则的每个条件赋一个有效度。有效度说明了该条件在特定情况下与决策属性的相关性。有效度越大，相关性就越大。删除条件的操作应先删除有效度最小的条件。

3 属性有效度量的选择

如果从信息论的观点来看，条件属性与决策属性之间的相关性也就是条件属性相对于决策属性的信息量，可以用它们之间的平均互信息反映出来。已有的最大泛化规则生成算法将属性的有效度定义为[1]：

$$SIG(c_i)=p(c_i)(p(D|c_i)-p(D)) \quad (1)$$

其中 $p(c_i)$是条件 c_i 的发生概率，$p(D|c_i)$是决策 D 关于条件 c_i 的条件概率，$p(D)$是决策 D 的发生概率。这种基于概率的度量，主要有两个缺点，一是有可能会出现负值，二是对较少出现的属性值赋予较小的有效度值，并不能真正反映出它的信息量。

这里，我们引入基于信息的 J-measure[2,3]作为属性有效度量。J-measure 曾被 Smyth P 和 Goodman R M 引入著名的规则归纳算法 ITRULE，作为规则的重要性度量，从而在一组数据实例中学习一系列优化的规则。而在最大泛化规则生成算法中引入 J-measure，是作为规则中每一个条件属性的有效度量。根据计算出来的有效度对一条规则的全部条件属性进行排序，以便按有效度从小到大的顺序尝试删除属性。

假设 D 是决策属性，它有 k 个值，记为 $d_1, d_2, \cdots, d_k$。条件属性记为 c_i。J-measure 定义为

$$J(D;c_i)=p(c_i)\sum_k p(d_k|c_i)\log\frac{p(d_k|c_i)}{p(d_k)}$$

因为一条规则给出的信息只是关于决策 D 及它的补 $\overline{D}$，所以以上的公式可以简化为：

$$J(D;c_i)=p(c_i)(p(D|c_i)\log\frac{p(D|c_i)}{p(D)}+(1-p(D|c_i))\log\frac{1-p(D|c_i)}{1-p(D)} \quad (2)$$

J-measure 作为一种信息度量，具有如下特点：1）显然，J-measure 具有非负性；2）满足$\sum_i J(D;c_i)=I(D;c)$，其中，$I(D;c)$就是 Shannon 定义的 D 和 c 之间的平均互信息。这个等式说明：一个属性所有值的信息量之和恰好是该属性与决策属性的平均互信息；3）与基于概率的有效度量不同，J-measure 采用对数计算，反映的是属性的信息量，受属性取值的频度影响较小。因此，J-measure 是一种比较理想的属性有效度量。

4 一个例子

这里，我们给出一个小型的数据库样本进行说明。这个样本是关于日本和美国汽车的数据[1]，我们希望从这些数据中发现一些有用的分类规则。先通过属性泛化处理生成一个泛化信息表，再计算该泛化信息表的约简，如表1所示。可以由约简直接写出一个规则集，将采用上述两种有效度量的算法分别用于约简，计算最大泛化规则。

表1 一个约简

make_model	weight	power	comp	trans	mileage
USA	medium	high	high	auto	medium
USA	medium	high	medium	manu	medium
USA	heavy	high	high	manu	medium
USA	medium	high	high	manu	medium
USA	light	high	high	manu	high
USA	medium	medium	medium	manu	medium
USA	heavy	high	medium	manu	low
Japan	light	low	high	manu	high
USA	medium	low	high	manu	medium
USA	medium	medium	high	auto	medium
Japan	medium	low	high	manu	high
Japan	light	medium	medium	manu	high
Japan	medium	high	high	manu	high
Japan	medium	low	medium	manu	high
USA	heavy	high	medium	auto	low
USA	medium	high	medium	auto	medium
Japan	medium	medium	high	manu	high
USA	medium	medium	high	manu	high

比如，表1中的第7项可以写成

if(make_model=USA)&(weight=heavy)&(power=high)&(comp=medium)&(trans=manu)
then(mileage=low)

若将属性有效度定义如(1)式，可得：1)SIG(tran=manu)=-0.03；2)SIG(make_model=USA)=0.04；3)SIG(power=high)=0.06；4)SIG(comp=medium)=0.07；5)SIG(weight=heavy)=0.093。

若将属性有效度定义如(2)式，可得：1)SIG(make_model=USA)=0.004；2)SIG(tran=manu)=-0.008；3)SIG(power=high)=0.018；4)SIG(comp=medium)=0.028；5)SIG(weight=heavy)=0.186。

其中的概率和条件概率计算均依据原始的数据库样本。可以看到，两种方法计算出的顺序并不相同。按上面第二种顺序删除条件。前三个条件删除后没有引起不一致的结果，但第四个条件"comp"删除后，新规则"if(weight=heavy)then(mileage=low)"与表1中的第3项所导出的规则不一致，因此，条件"comp"被恢复。同样，第5个条件"weight"不能删除。所以，由这条具体规则导出的最大泛化规则是：

（下转第113页）

结点、表示客体、客体的性质、概念、事件、行为等，有向边表示结点间的关系，语义网络知识库的刷新为有向图中"结点-连接边"的增减或修改。

2. 框架表示[5] 框架是由 Minsky1975年提出的，本意是将语义网络过程等知识表达方式结合起来描述一些固定的环境和常规行为，一般将它归类为一种聚集语义网[5]。其中的语义网聚集结点对应于框架名，其聚集元素为框架的槽(slot)，这些槽及其相应的值构成了框架的一个联合表达。因此，根据联合最大熵定理，在使用框架表达方式时，应使其各个槽及相应的值相互独立，这个结点基本与人们的实际使用经验相符合。

3. 逻辑表达 逻辑是由一组严格定义的表达知识的语法符号，加上一些严格定义的对这些符号表达的知识进行解释和操作的形式化方法组成，一般说来，任何满足以下条件的知识处理系统即可看作是一个逻辑"系统"[1]。

(1)有一种严格定义的知识表达语言(语法结构)；

(2)有一种严格定义的模型理论(语义)，给这种语言的语句赋值的语义；

(3)有一种严格定义的证明理论及形式推理方法，对这种语句进行语法操作，从一些语句推导出另一些语句。

4. 本文基于 Agent 的知识表达方法和前述的方法相比主要区别有以下几点：(1)Agent 它本身是一个智能体，它自己有推测、反响、自学习和相互学习等能力；(2)Agent 它本身就可以作为一个知识库或信息库来处理，贮存知识；(3)用 Agent 表示知识真正从本质上解决了智能化问题。

结论 本文抓住 Agent 的特点，研究 Agent 的知识表达度量理论，归纳起来作出如下贡献：

1. 给出了 Agent 知识量的概念，为进一步研究 Agent 的度量理论奠定了基础；

2. 利用熵这个概念，对 Agent 的知识进行量化处理，找到了 Agent、Agent 与 Agent 之间的量化理论依据；

3. 给出了联合表达最大熵定理，复合熵增定理，联合熵增定理等基本理论。

通过本文的研究，进一步丰富和发展了 Agent 系统理论的研究内容，为 Agent 走向实用化、商品化打下坚实的理论基础。

参考文献

1 Hu SHanli, et al. The logic tool used in the formalized research on agent. Computer Science, 1999, 26(12): 1～4

2 Davis R, Shroke H, Szlovits P. What is a knowledge representation? AI Magazine, 1993, spring: 17～33

3 Fu-Tong, et al. Knonledge Engineering. Beijing, Science Press, 1992

4 Dietterech T G. Exploratory research in machine learning. Machine Learning, 1990(5): 5～9

5 Li Fanzhang, et al. Principle and method and applications of artificial intelligence. Kuming, Yunnan Science Press, 1997

(上接第115页)

if (weight = heavy) & (comp = medium) then (mileage=low)

表2是表1所导出的最大泛化规则集。

表2 一组最大泛化规则

make_model	weight	power	comp	trans	mileage
-	heavy	-	medium	-	low
USA	medium	high	-	-	medium
USA	medium	-	medium	-	medium
-	medium	-	-	auto	medium
USA	-	light	-	-	medium
-	heavy	-	high	-	medium
-	-	medium	high	manu	high
Japan	-	-	-	-	high
-	light	-	-	-	high

小结 发现分类规则是分类知识发现的核心任务。在基于 RS 理论的发现方法中，可以从信息表的约简直接写出规则，并进一步求出最大泛化规则。在求最大泛化规则的过程中，尝试删除条件属性，在不降低规则的分类精确度的前提下使规则长度尽可能地短，而删除条件属性的顺序由条件的属性有效度来决定。本文引入 J-measure 作为属性的有效度量，在开始删除操作前计算出属性的有效度并排序，可以生成较优化的最大泛化规则。

参考文献

1 Cercone N, Hamilton H, Hu X, Shan N. Data Mining Using Attribute-Oriented Generalization and Information Reduction. In: Rough Sets and Data Mining - Analysis of Imprecise Data Mining Eds. Lin T Y and Cercone N Kluwer Academic Publishers, Boston/London/Dordrecht, 1997. 199～227

2 Smyth P, Goodman R M. Rule Induction Using Information Theory. In: Knowledge Discovery in Databases. Eds. Piatetsky-Shapiro G and Frawley W J. Cambridge, MA: AAAI/MIT Press, 1991 159～176

3 Smyth P, Goodman R M. An Information Theoretic Approach to Rule Induction from Databases. IEEE Transactions on Knowledge and Data Engineering, 1992, 4(4): 301～316

4 徐如燕. 基于粗糙集理论的知识发现研究:[硕士学位论文] 空军雷达学院, 2000