内涵缩减与分类规则求解

葛 斌[1],孟祥瑞[2]

1. 安徽理工大学计算机科学与工程学院,安徽淮南 232001
2. 安徽理工大学科研处,安徽淮南 232001

摘要 概念格是数据分析与知识提取的一种有效工具,具有精确性和完备性等特点。目前,基于概念格的分类规则提取算法很多,但在提取到规则的数量上和规则的形式上并不能达到令人满意的效果。针对基于概念格的分类规则提取方法进行了研究,在改进内涵缩减的增量式计算方法基础上给出了基于内涵缩减的确定的分类规则和近似的分类规则的提取方法,通过有效限制计算内涵缩减的节点的范围降低了内涵缩减的计算规模,利用分类规则基,降低了需要计算的分类规则的数量,提高了分类规则的提取效率。为验证本研究提出分类关联规则的挖掘方法,用 C++实现了上述算法。测试结果表明,本文给出的算法是有效的。

关键词 形式背景;概念格;内涵缩减;分类规则

中图分类号 TP18 **文献标识码** A **文章编号** 1000-7857-(2009)15-0071-05

Intent Reduction and Classification Rules Mining

GE Bin[1], MENG Xiangrui[2]

1. *College of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 23200, Anhui Province, China*
2. *Department of Scientific Research, Anhui University of Science and Technology, Huainan 232001, Anhui Province China*

Abstract The concept lattice is accurate and complete in the knowledge representation, and it is an effective tool for data analysis and knowledge discovery. Classification rule mining based on the concept lattice in an efficient way is a challenging task. At present, there are many extraction algorithms for classification rule based on the concept lattice, but the number and form of extracted rule can not achieve satisfactory results. This paper focused on classification rule mining using intent reduction of the formal concept. It proposed an incremental way to compute intent reduction by adding object to the formal context one by one. Through modifying the incremental computation of the intent reduction of concepts, new algorithms are developed to compute the intent reduction, and then definitions of the exact classification rule base and the approximate base are given. On the basis, two algorithms for the exact and the approximate mining bases are designed. In the algorithms, only those concepts that concern the classification rule mining need to be considered in compute their intent reduction. Furthermore, the use of classification rule bases reduces the total number of classification rules that need to be extracted. In order to verify the data mining methods of classification association rules which were put forward in this research, the algorithms mentioned above were implemented by using C++. Finally, empirical experiments on UCI data demonstrated the efficiency of these algorithms in the mining of affirmative classification rules and approximate classification rules using intent reduction.

Keywords formal context; concept lattice; intent reduction; classification rule

基于概念格的分类规则提取算法很多,但在提取到规则的数量上和规则的形式上并不能达到令人满意的效果。例如,Godin 等[1]、王志海等[2]分别提出了由概念格来提取蕴涵规则的算法,但是这些规则大多是确定性规则,而实际数据库往往含有大量的错误或不确定的数据,确定性规则往往不存在或数量很少。而另外一些算法,会产生大量高可信度的冗余规则。本文借鉴文献[3]的思想,用确定的分类规则基和近似的分类规则基进行分类,提出了基于同步剪枝和同步内涵缩减计算的分类规则提取方法提取这两类分类规则,在不丢失有用的信息的前提下,有效减少了分类规则的数量。

收稿日期:2009-04-10

基金项目:安徽高校省级自然科学研究重点项目(KJ2009A005Z)

作者简介:葛斌,研究方向为系统工程及计算机应用,电子信箱:bge@aust.edu.cn;孟祥瑞(通信作者,中国科协所属全国学会个人会员登记号:E382370003M),教授,研究方向为采矿工程、系统工程,电子信箱:xrmeng@aust.edu.cn

1 相关定义

首先给出应用内涵缩减提取分类规则[4]的相关定义。

定义1 确定的分类规则基为

$$\mathrm{ER}\ B=\{r:red\Rightarrow c|c\in B\wedge red\in \mathrm{IRed}(B)\wedge c\notin red\}$$

其中,B为包含分类属性c的频繁封闭项集(对应频繁概念的内涵),$\mathrm{IRed}(B)$为B的缩减集合。

确定的分类规则$r:red\Rightarrow c$的置信度$\mathrm{Conf}(r)=1$,即$\mathrm{Sup}(red\cup\{c\})=\mathrm{Sup}(red)$,这意味着前件中的$red$和$c$出现于同一概念的内涵中,因此,要提取确定的分类规则只需考查包含当前类别c的格节点。

定义2 近似的分类规则基为

$$\mathrm{ARB}=\{r:red\Rightarrow c|c\in B\wedge red\in \mathrm{RED}\wedge \mathrm{f}(\mathrm{g}(red))\subset B\}$$

其中,B为包含当前类别c的频繁封闭项集(即频繁概念的内涵),RED为所有频繁封闭项集的缩减集合。

近似的分类规则$r:red\rightarrow c$的置信度$\mathrm{Conf}(r)<1$,即$\mathrm{Sup}(red\cup\{c\})<\mathrm{Sup}(red)$,意味着包含前件$red$的节点应为包含类别$c$的节点的前驱节点,而包含类别$c$的前驱节点在定义1确定的分类规则已讨论过,因此对近似的分类规则只要考虑不包含类别c的前驱节点。

文献[3]指出,所有有效的确定分类规则,包括它的支持度和可信度都可以从确定的分类关联基及其支持度导出,因此只要根据定义利用内涵缩减找到满足用户指定的最小支持度和最小可信度的所有确定的分类规则基和近似的分类规则基,再构造出分类器即可。

要提取确定的分类规则,只需考查包含当前类的格节点;要提取近似的分类规则,只需考查包含当前类的格节点与其前驱节点间的关系。因此本文考虑:生成包括当前类的节点及其前驱节点的一个部分格,并在此基础上提取分类规则。

2 改进的内涵缩减同步计算方法

文献[5]提出基于增量式建格方法的内涵缩减求解,本文考虑在其方法基础上,完成确定的分类规则基和近似的分类规则基的提取。因为只需要2类节点的内涵缩减:第1类是满足阈值条件的并且含有类别c的节点,第2类是满足阈值条件、不含有类别c而且有第1类节点是它的超概念的节点。在增量式内涵缩减求解过程中,可能会出现冗余计算,为完成分类规则的提取,需要进一步改进算法,提高算法效率,研究这2类节点在增量过程中的变化很有必要。

设形式背景为(G,M,I),记$MinSup$为最小支持度阈值,L为其对应的概念格,$\tilde{L}$为对应的剪枝部分格,是将L中外延基数小于最小支持度阈值$MinSup$的节点删除后的部分,它是L的过滤,也是上确界子格;假设现将属性m加入该背景,$G_m=\mathrm{g}(m)\subseteq G$是具有属性$m$的对象集合,即新形势背景为$(G,M\cup\{m\},I\cup G_m\times\{m\})$,设该新背景对应的概念格为$L^*$,$\tilde{L}^*$为在$\tilde{L}$中插入属性$m$后的部分格,$\overline{L}^*$为$\tilde{L}^*$对应的剪枝部分格。

定理1 $(A,B)\in\tilde{L}$,且$(A,B)\notin\tilde{L}_1\cup\tilde{L}_2$,这里$\tilde{L}_1$为$\tilde{L}$中第1类节点集合,$\tilde{L}_2$为$\tilde{L}$中第2类节点集合,那么当$(A,B)$更新后对$\tilde{L}_1$和$\tilde{L}_2$中节点的缩减没有影响。

证明 只要证明(A,B)更新后$\tilde{L}_1$和$\tilde{L}_2$中节点的后代不发生变化即可。分两种情况讨论:如果(A,B)为更新节点,结论是平凡的;如果(A,B)为生成元,记$nCon$为(A,B)对应的新节点,且满足阈值条件,显然$(A,B)\in\tilde{L}^*$,此时如果存在$pCon$是其在$\tilde{L}^*$中的父节点,有2种情况:$pCon$也是新节点,那么$pCon$不可能属于集合$\tilde{L}_1$,因为$pCon$的外延只比A多了x,同样$pCon$不可能属于集合$\tilde{L}_2$,否则由于$pCon$的超概念也是(A,B)的超概念,从而有(A,B)为第2类节点,这与$(A,B)\notin\tilde{L}_1\cup\tilde{L}_2$矛盾,故结论成立;$pCon$是更新节点,类似前者可知$pCon$不属于$\tilde{L}_1\cup\tilde{L}_2$,从而结论成立。

由定理1可知,在$\tilde{L}$中不属于$\tilde{L}_1\cup\tilde{L}_2$的节点在更新后对$\tilde{L}_1\cup\tilde{L}_2$中节点缩减的计算没有影响。如果$\tilde{L}_1$中节点是满足阈值条件的生成元或更新节点,容易知道对应的新节点和更新后的节点仍为第1类节点;如果$\tilde{L}_2$中节点是满足阈值条件的生成元,则需要根据其父节点的情况判定,如果其对应的新节点不是第2类节点,那么由定理1知道如果该新节点可以不参加下一步的增量过程,即不属于更新后的剪枝格。这样就可以将$\tilde{L}$中不属于$\tilde{L}_1\cup\tilde{L}_2$的节点直接删除,不参与增量过程,此时记$\tilde{L}:=\tilde{L}_1\cup\tilde{L}_2$,并且记$\tilde{L}^*:=\tilde{L}_1^*\cup\tilde{L}_2^*$,即同步剪枝后得到的仍都是第1类和第2类中的节点,这里仍考虑对偶情况,算法如下:改进的基于增量式建格内涵缩减同步计算MIIR,图1为算法MIIR流程图,图2为内涵缩减算法Compute流程图。

函数JUDGE(Con)在商半格中判断Con是否拥有超概念是第一类节点,如果是则返回真值。通过JUDGE的返回决定某个生成元是否需要生成新节点,从而降低下一轮增量过程的计算量。

按照上面改进的同步剪枝算法可以得到只包含第1类和第2类节点的部分格,它仍然是概念格L^*的一个理想,因此其节点缩减的计算是不发生变化的,上面的算法MIIR实现了所需节点内涵缩减的同步计算。

3 基于内涵缩减的分类规则提取

利用上面得到的剪枝格(每个节点的缩减存储在其$ERed$域),根据定义1和定义2提取2类分类规则。给出基于剪枝格的分类规则基提取算法CRPL(流程图见图3),首先调用上面改进的基于增量式建格的内涵缩减计算方法MIIR得到只包含第1类和第2类节点的部分格以及相应的缩减集,随后用$T1$保存所有的第1类节点,用$T2$保存$T1$中存在父节点属于第2类节点的边界节点,然后先对$T1$调用ERB(流程图见图4)提

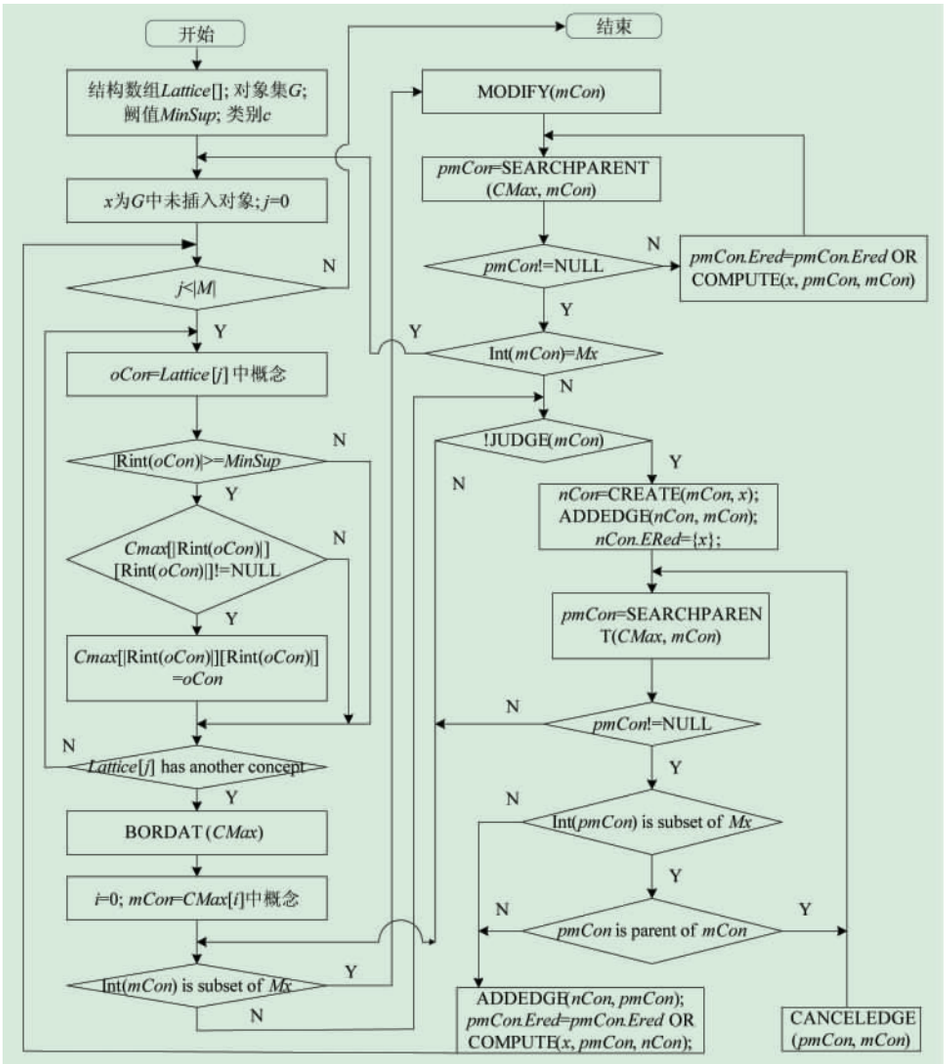

图 1　MIIR 算法流程

Fig. 1　Flow chart of algorithm MIIR

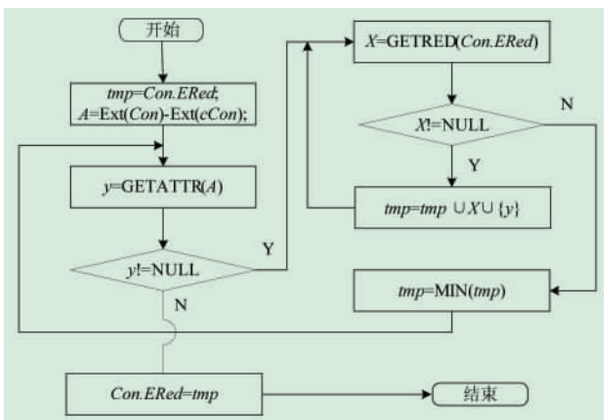

图 2　Compute 算法流程

Fig. 2　Flow chart of algorithm Compute

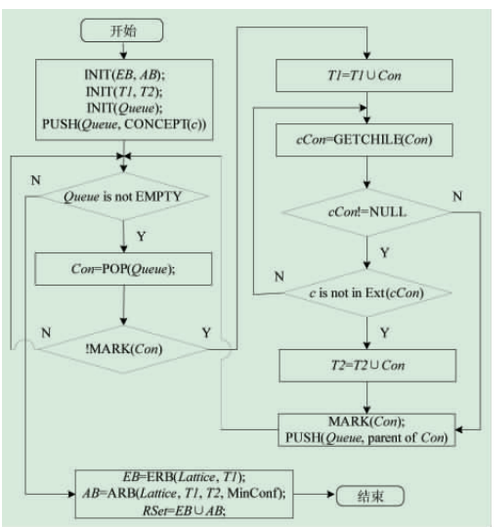

图 3　CRPL 算法流程

Fig. 3　Flow chart of algorithm CRPL

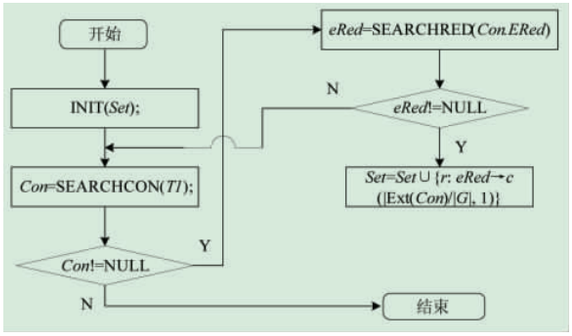

图 4　ERB 算法流程

Fig. 4　Flow chart of algorithm ERB

取所有确定的分类规则基,再对 $T1$、$T2$ 调用 ARB(流程图见图 5)提取所有近似的分类规则基,从而得到所需的分类规则集。

基于剪枝格的分类规则基提取算法 CRPL 流程图见图 3。CRPL 输入/输出为:输入:$Lattice[\ ]$, G,类别 c,支持度阈值 $MinSup$,置信度阈值 $MinConf$;输出:分类规则集 $Rset$。

4　实验分析

为验证本文提出分类关联规则的挖掘方法，用 C++实现了上述算法并使用文献[6]提出的启发式方法对前面得到分类关联规则集构造分类器 CSCR 并在 UCI 机器学习库中的 6

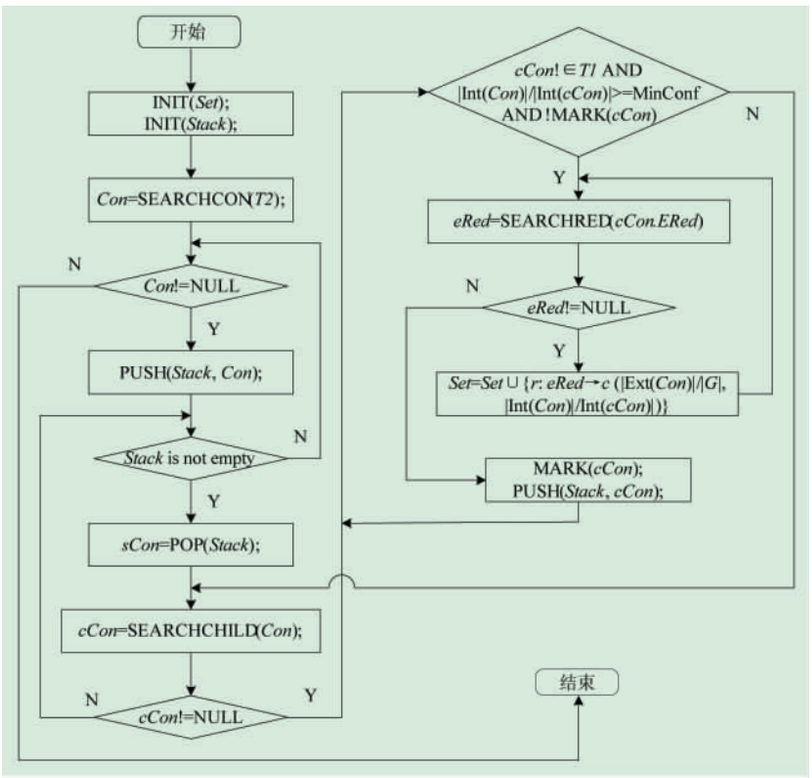

图 5　ARB 算法流程

Fig. 5　Flow chart of algorithm ARB

个数据集上进行测试,其中数据集信息如表 1 所示。

试验采用四交叉验证技术。最小支持度的取值对最终的分类器精度影响很大,实验过程中发现当最小支持度取值在 1%左右时获得的规则数适中,分类器的效果最好。因此,在最小支持度为 1%、最小置信度为 60%的情况下对上面 6 个数据集进行了和标准分类方法 C4.5 的对比实验，结果见表 2,表 2 中第 1 列为数据集,第 2 列为 C4.5 的分类精度,第 3 列为 Classifier 算法的分类精度,第 4 列为本文算法提取分类关联规则的时间,第 5 列为提取的分类关联规则个数。

从表 2 可以看出,CSCR 算法在参与测试的 6 个数据集里有 4 个分类精度高于 C4.5,尤其在 glass 数据集上比 C4.5 有较大的提高,在 2 个数据集上分类精度低于 C4.5。在 6 个数据集上的分类精度均值分别是:CSCR 为 83.8%,C4.5 为 81.58%,即本研究提出的分类方法优于标准 C4.5 方法。另外,从 CSCR 提取分类关联规则的所需时间看,有 4 个数据集时间较快,平均 1.4 s 的时间是可以接受的。而 breast-w 和 heart 2 个数据集由于提取的规则数较多因此需要的时间略长,尽管如此,不到 8 s 的平均时间和文献[6]中相比速度上仍是有所提高。

表 1　实验数据集

Table 1　Data sets used in experiments

数据集	属性数	类别数	对象数
breast-w	10	2	699
diabetes	8	2	768
glass	9	7	214
heart	13	2	270
iris	4	3	150
pima	8	2	768

表 2　CSCR 与 C4.5 的对比结果

Table 2　Results of algorithm CSCR and C4.5

数据集	C4.5	CSCR	时间	规则数
breast-w	95.0	96.3	28.7	307
diabetes	74.2	73.1	2.6	35
glass	68.7	81.1	1.3	49
heart	80.8	83.2	10.9	528
iris	95.3	96.1	0.4	19
pima	75.5	73.0	1.2	25
平均值	81.58	83.8	7.52	160.5

5　结论

提出以增量式建格方法为基础、基于内涵缩减同步计算的分类关联规则挖掘方法是有效的。本文算法与文献[3]相比有两方面改进：限制了计算内涵缩减的节点范围，只需计算和分类规则提取有关的节点的内涵缩减；在分类规则提取时由于所需的内涵缩减已经增量式计算，因此两类分类规则的提取可以直接根据前面的定义给出，提高了规则提取效率。但是，由于本文测试的数据集还不是很多，不同的数据集对应的训练背景可能差异很大，因此未来的工作之一就是测试更多的数据集以进一步完善本文提出的算法。

参考文献(References)

[1] Godin R, Missaoui R. An incremental concept formation approach for learning from databases [C]//Theoretical Computer Science, Special Issue on Formal Methods in Databases and Software Engineering. October 1994, 133: 387-419.

[2] 王志海, 胡可云, 胡学钢, 等. 概念格上规则提取的一般算法与渐进式算法[J]. 计算机学报, 1999, 22(1): 66-70.
Wang Zhihai, Hu Keyun, Hu Xuegang, *et al*. *Chinese Journal of Computers*, 1999, 22(1): 66-70.

[3] 齐红, 刘大有, 刘亚波. 基于概念格的用户关联挖掘 [J]. 计算机科学, 2004, 31(Z2): 158-161.
Qi Hong, Liu Dayou, Liu Yabo. *Computer Science*, 2004, 31 (Z2): 158-161.

[4] 谢志鹏, 刘宗田. 概念格与关联规则发现[J]. 计算机研究与发展, 2000, 37(12): 1415-1421.
Xie Zhipeng, Liu Zongtian. *Journal of Computer Research and Development*, 2000, 37(12): 1415-1421.

[5] Liu B, Hsu W, Ma Y. Integrating classification and association rule mining [C]//Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. New York, 1998: 80-86.

[6] 胡可云, 陆玉昌, 石纯一. 基于概念格的分类和关联规则的集成挖掘方法[J]. 软件学报, 2000, 11(11): 1478-1484.
Hu Keyun, Lu Yuchang, Shi Chunyi. *Journal of Software*, 2000, 11(11): 1478-1484.

(责任编辑　杨书卷)