文章编号:1001-9081(2007)10-2376-03

基于扩展概念格的分类规则获取算法

王 燕,李 明
(兰州理工大学 计算机与通信学院,兰州 730050)
(wangyan@sohu.com)

摘 要:概念格是进行数据挖掘和规则提取的有力工具,通过分析概念格中概念的特征,提出了扩展概念格以及基于扩展概念格的分类规则获取算法。实验表明该算法能够生成简洁并且易于理解的规则集。

关键词:概念格;扩展概念格;分类规则获取;新颖性
中图分类号:TP311.131 **文献标志码**:A

Classification rule acquisition based on extended concept lattice

WANG Yan, LI Ming
(College of Computer and Communication, Lanzhou University of Technology, Lanzhou Gansu 730050, China)

Abstract: Concept lattice is a powerful tool for data mining and rule acquisition. Through analyzing the characteristics of concept in concept lattice, extended concept lattice and classification rule acquisition based on extended concept lattice was proposed. Experimental results show that this algorithm could obtain simple and understandable rule set.

Key words: concept lattice; extended concept lattice; classification rule acquisition; novelty

0 引言

概念格,也称为 Galois格,又叫做形式概念分析,首先由 Wille R于 1982年提出[1]。概念格的每个节点是一个形式概念,由两部分组成:外延,即概念所覆盖的实例;内涵,即概念的描述,也叫该概念覆盖实例的共同特征。概念格通过 Hasse图生动而简洁地体现了这些概念之间的泛化和特化关系。因此。概念格被认为是进行数据分析的有力工具。目前,已有一些建造概念格的算法和概念格在信息检索、数字图书馆、软件工程和知识发现等方面的应用[2-5]。分类规则获取是重要的数据挖掘任务。典型的分类规则获取方法有决策树、神经网络以及粗糙集等。实验评估显示,基于概念格的系统具有和这些典型系统同样良好的分类效果[3,4,6-7]。

基于概念格的知识获取主要包括获取关联规则和分类规则[8]。现有的基于概念格获取分类规则的算法主要采用两种策略:一种是利用建立概念格的算法,建立一个完全格,如 GALOIS[9]算法和 RULEARNER[10]算法。这两种算法都采用增量式方法产生数据集所对应的一个完全格,然后进行规则提取。利用这种方法提取的规则不能描述数据中包含的不确定性,也就是说,当数据中有噪声存在时,算法的正确识别率将会受到影响。另外一种策略是在一定的学习参数的控制下,建立数据集所对应的一个半格,如 LEGAL[11]算法、LACS[12]算法和 CBALattice[13]算法。这三个算法虽然可以产生概率规则,具有抗噪声的能力,但算法中所涉及到的学习参数都需要人为设定。如果人们对待研究的问题还没有很好的认识,设定合适的、有利于系统性能的学习参数将是很困难的。另外,不论通过哪种策略提取分类规则,每一个规则都是由概念格中的概念内涵生成的。由概念格的特性可以知道,概念内涵是概念外延对应的最大数据集,由它生成的规则是比较繁琐的。

因此,本文通过概念格中概念的特征,提出了扩展概念格以及基于扩展概念格的分类规则获取算法,并在算法中利用提出的规则新颖性对产生的规则集进行剪枝,从而可以获得简洁而容易理解的规则。实验表明算法是可行并且有效的。

1 基本概念[14]

定义 1 给定三元组 $K=(U, A, R)$为一个形式背景,其中 U为对象集合;A为属性集合;R为 U和 A之间的二元关系,即 $R\subseteq U\times A$,对于 $x\in U$, $a\in A$, xRa读作“对象 x具有特征 a”。形式背景 K中,在 P(U)和 P(A)上定义了两个映射关系 f和 g:

$$\forall X\subseteq U: f(X)=\{a\in A \mid xRa, \forall x\in X\}$$

$$\forall B\subseteq A: g(B)=\{x\in U \mid xRa, \forall a\in B\}$$

通过这两个映射关系,能够定义形式概念及概念格。

定义 2 给定形式背景 K上的一个序偶 (X, S),其中 $X\subseteq U$, $S\subseteq A$,如果满足 $f(X)=S$并且 $X=g(S)$,则称 (X, S)为一个概念。X称为概念 (X, S)的外延,S称为概念 (X, S)的内涵。形式背景 K中的所有概念及概念之间的偏序关系构成的结构称为概念格,记作 L(U, A, R)。其中每个概念被看作概念格中的一个节点。

由上述的定义可以知道,概念的内涵是该概念外延对应的最大属性集。而对于一个形式背景,还存在着另外一些由对象和属性构成的二元组,虽然它们不满足定义 1中的两个映射关系,但是这样的二元组能够简化概念格中概念的内涵,我们称其为扩展概念。定义如下:

定义 3 给定形式背景 K上的一个序偶 (X, S),其中 $X\subseteq U$, $S\subseteq A$,如果该二元组满足 $X=g(S)$并且 $S\subseteq f(X)$,则称 (X, S)是形式背景上 K的扩展概念。

定理 1 给定形式背景 K(U, A, R), (X, S)为形式背景上的扩张概念。那么在形式背景上肯定存在一个概念 (X, T),使

收稿日期:2007-04-05;**修回日期**:2007-06-22。

作者简介:王燕 (1971-),女,甘肃泾川人,讲师,博士研究生,主要研究方向:数据挖掘、机器学习等; 李明 (1959-),男,甘肃兰州人,教授,主要研究方向:数据挖掘、知识工程、智能检索技术、图像处理等。

得 S⊆ T。

证明　由概念的定义可知,概念的内涵是该概念外延对应的最大属性集。由于 (X, S) 是一个扩展概念,则 S并不是对象集合 X所对应的最大属性集。那么肯定存在另一属性集 T,使得 X = g(T)并且 T = f(X),则 (X, T)就是形式背景上的一个概念。根据定义 3, X = g(S)并且 S⊆ F(X),则 S⊆ T。

根据上面的描述可知,形式背景的扩展概念与某个概念覆盖了相同的对象,而内涵是该概念的子集。所以用扩展概念生成的规则将更加简洁,并且可以证明其生成的规则可信度与利用概念生成的规则可信度是一样的。

2　基于扩展概念格的分类规则获取算法

LACS算法是一个基于概念格的处理多类分类问题的分类规则算法。算法通过概念纯度处理数据中的不确定性,并结合概念强度控制概念格的生成和有效地进行规则剪枝。算法中涉及到的参数需要由用户给出,而且规则由概念生成,使得获得的规则比较繁琐。基于扩展概念格的分类规则获取算法 (CAECL)对 LACS算法进行了以下改进: 1)用文献 [15]提出的数据集的局部最小确定性 α_c作为 LACS中概念纯度的阈值来处理数据的不确定性。在算法 LACS中,概念纯度表示了数据集的不确定性,通过对概念纯度的度量,如果概念纯度的值大于某个用户给定的阈值,则视为候选规则,反之,则舍弃该规则。但是在不同的应用领域中,用户很难设置合适的不确定性阈值。因此,我们选择 α_c作为概念纯度的阈值,因为 α_c是由数据集中的数据计算得到的,不需要用户设定。另外,在文献 [15]中已经证明,局部最小确定性 α_c反映了数据的不确定性,并且能够作为阈值来控制不确定规则的生成。2)用规则的新颖性代替概念强度对规则进行剪枝。概念格的建立是一个非常费时的工作,如果对一个数据集建立一个完全格,一是效率比较低,其次由该完全格获得的规则将会产生大量的冗余。因此,算法 LACS采用概念强度对规则进行剪枝,也就是删除不满足概念强度阈值的规则。在删除的过程中,只是根据用户给定的阈值,如果用户给出的阈值不合理,将会删除有用的规则。而下面定义的规则的新颖性,能够客观的反映一条规则是否冗余,是否被规则集中的规则所覆盖。利用规则的新颖性既可以达到删除冗余规则的目的,也可以不受用户主观的影响而得到给有用的规则。 3)在建立概念格时,我们采用自顶向下的原则搜索数据中的概念或扩展概念,这样可以使得获得的规则更加简洁。

一般分类规则都是从决策表中获取的。决策表通常是一个多值系统,也就是说,决策表中的每个属性对应着多个属性值,而形式背景是二值的。因此,在算法进行以前,需要把决策表转换为形式背景。具体作法是:把决策表中的每一个条件属性和决策属性的取值,当作形式背景的一个属性,并且把决策表的条件部分与决策部分分别转换为条件形式背景和决策形式背景,分别记作 (U, ConCorrespond(C), R_C) 和 (U, DesCorrespond(D), R_D),其中, U是对象的集合, C和 D分别是决策表的条件属性集和决策属性集, ConConversion(C) 和 DesConversion(D) 分别是条件属性和决策属性经过转换得到的形式背景的新属性。下面先给出规则新颖性的定义:

定义 4　给定条件形式背景 (U, ConCorrespond(C), R_C) 和决策形式背景 (U, DesCorrespond(D), R_D)。ObjectSet是已获得的规则集中规则覆盖的对象集合。对于一条新的规则 S→ T, S⊆ ConCorrespond(C)并且 T⊆ DesCorrespond(D),如果 g(S)⊄ ObjectSet,则称规则 S→ T是新颖的,反之称规则 S→ T是冗余的。

根据规则新颖性的定义,如果新生成的规则能够被规则集中的另一条规则所蕴涵,那么新得到的规则将被认为是冗余的,而不被写入规则集合中。这不仅可以简化规则集,同时能够提高算法的效率。

基于以上的描述和定义,我们的算法按照自顶向下分层搜索的原则寻找满足条件的概念或扩展概念。在每一层中,概念或扩展概念用四元组 (X, S, Y, T)表示, X, Y是对象的集合, S⊆ ConConversion(C), T⊆ ConConversion(D),并且四元组满足 X = g(S), S⊆ f(X), Y = g(T), T = f(Y),对于规则的概念纯度大于 α_c的概念或扩展概念,可以生成规则;否则,概念或扩展概念被忽略而不生成规则。每一层结束通过判断目前的规则集中规则覆盖的对象数是否达到了整个对象集,来预测下一层的概念或扩展概念是否会生成新颖的规则。如果不能通过预测则算法结束,否则进入下一层。由于算法中的阈值 α_c是由数据计算出来的,不需要用户参与从而可以实现自主的规则获取;同时对规则新颖性的判定,可以减少概念的搜索空间,提高算法的效率。在下面的算法描述中,我们把概念和扩展概念统称为概念。

算法:　基于扩展概念格的分类规则获取算法 CAECL

输入:决策表 S = ⟨U, R, V, f⟩,条件形式背景 (U, ConCorrespond(C), R_C) 和决策形式背景 (U, DesCorrespond(D), R_D)。RuleSet = ϕ为规则集。

输出:分类规则集 RuleSet。

第 1步　根据文献 [15]中的定义计算决策表局部最小确定性 a_c,以 a_c作为控制规则生成的阈值。

第 2步　初始化最顶层的概念: L_1 = {(g(S), S, g(T), T)},其中, g(S), g(T) ⊂ U, S⊂ ConConversion(C), T⊂ DesConversion(D), (g(S), S)和 (g(T), T)分别是条件形式背景上的概念和决策形式背景上的概念。并且满足 $g(S) = \max_{M\subset ConConversion(D)}\left(\frac{|g(S)\cap g(M)|}{|g(S)|}\right)$。对 L_1中的概念,如果 $\frac{|G(S)\cap G(T)|}{|G(S)|} > \alpha_c$,则生成规则 S→ T, RuleSet = RuleSet ∪ {S→ T}。

第 3步　建立下层概念格,并获得规则。

1)预测由下层概念生成的规则是否具有新颖性,如果有,则继续 2),否则,算法结束。

2)当 L_i层中的概念个数大于 1时,设 L_{i+1} = ϕ。对 L_i中所有的概念对 (g(S), S, g(T), T)和 (g(S′), S′, g(T′), T′)进行合并运算: g(Z) = g(S)∩ g(S′)。

3)对每一个合并,当 g(Z)≠ ϕ时,则新概念的条件部分为 (g(Z), S∪ S′),决策部分取 |g(T)|和 |g(T′)|中较大的对象集合以及其对应的属性。假设大的为 |g(T)|,则决策部分为 (g(T), T)。L_{i+1} = L_i∪ (g(Z), S∪ S′, g(T), T)。

4)如果 |g(Z)| ≠ |g(S)|, |g(Z)| ≠ |g(S′)|,并且 $\frac{|g(Z)\cap g(T)|}{|G(Z)|} > \alpha_c$,则生成规则 S∪ S′→ T, RuleSet = RuleSet∪ {s∪ S′→ T}。

5)如果 |g(Z)| = |g(S)|或者 |g(Z)| = |g(S′)|,则不生成规则,继续下一个合并。

6)如果对 L_i层概念的两两合并全部处理完,则转入 2)。

3　实验结果

为了评估算法的有效性,我们使用了 UCI数据集[16]对算法 LACS和 CAECL进行测试。具体的实验方法为:从数据集

中随机删除一些属性,以增加数据的不确定性;将缺省属性值补齐为该属性中最频繁的属性值;将连续属性集利用“基于属性重要性离散化算法”[17]进行离散化处理;随机选择一定数量的数据作为训练集,其余的作为测试集;并采用高信任度优先法的规则选取策略消解冲突。在算法中采用的数据集见表 1。

表 1 数据集

数据集名	总样本数	训练样本数	测试样本数
Breast-Cancer	699	349	350
Balance-Scale	625	312	313
Letter	20000	2000	18000
Mushroom	8124	4012	4012
Car	1728	691	1037
Iris	150	75	75
Solar-flare	1066	533	533
Tic-Tac-Toe	958	479	479

我们在上面的数据集上运行算法 LACS,其中概念纯度从 0.1开始取值,步长为 0.1,概念强度从 0.01开始取值,步长为 0.01。并且在以上的数据集上运行算法 CAECL,用 α_c 作为概念纯度的阈值。从 LACS对每个数据集的测试结果中,挑出正确识别率最大并且规则较少的数据,和算法 CAECL的测试结果进行比较,如表 2所示。

从表 2可以看出,不同的数据集上,既要保证较高的正确率又要使规则数较少的概念纯度和概念强度的值是不同的。要由用户设定合适的参数值是比较困难的,并且采用从概念获取的规则集中规则的平均长度比较长。而采用 CAECL算法得到的结果,虽然不能完全达到最大正确率,但也是在最大正确率附近,并且规则数相对比较少,规则集中,规则的平均长度也较短。因此,由上面的实验数据表明,这种通过取决策表最小确定性和规则新颖在扩展概念格上获取不确定规则的方法是有效的。

表 2 算法 LACS和 CAECL的比较

数据集	LACS					CAECL			
	概念纯度	概念强度	正确率	规则数	规则平均长度	α_c	正确率	规则数	规则平均长度
Breast-Cancer	0.8	0.02	92.8	66	2.08	0.5	92.3	39	1
Balance-Scale	0.5	0.03	78.3	63	2.01	0.33	73.1	21	1
Letter-Recognition	0.1	0.01	40.1	63	1	0.2	40.1	40	1
Mushroom	0.8	0.01	60	33	3.57	0.65	74.5	10	1
Car	0.3	0.02	64	9	1	0.6	64	6	1
Iris	0.6	0.02	90.7	10	1	0.5	89.3	12	1
Solar-flare	0.7	0.04	97.6	40	2.59	0.5	97.4	21	1
Tic-Tac-Toe	0.6	0.02	72	43	2.89	0.5	71	9	1

4 结语

本文提出了一种基于扩展概念格的分类规则获取算法。该算法中涉及到的学习参数,不需要用户设定,完全通过训练数据计算得到;通过对规则新颖性的判断,能够删除冗余规则,提高算法效率;通过从扩展概念获得规则,缩短了规则集中规则的平均长度,使获得的规则更加的简洁。仿真实验结果表明该算法对获取分类规则是一种有效、可行的方法。

参考文献:

[1] WILLE R. Restructuring lattice theory: an approach based on hierarchies of concepts[M]. Dordrecht-Boston: Reidel, 1982. 445—470.

[2] GODIN R. Incremental concept formation algorithm based on Galois (concept) lattices[J]. Computational Intelligence, 1995, 11(2): 246—267.

[3] NJIWOUA P, NGUIFO E M. Forwarding thechoice of bias LEGAL-F: using feature selection to reduce the complexity of LEGAL[C]// Proceedings of the BENELEARN-97, ILK and INFOLAB. Netherlands: Tiburg University Press, 1997: 89—98.

[4] CARPINETO C, ROMANO G. Galois: anorder theoretic approach to conceptual clustering[C]// Proceedings of the ICML-93. Amhers: Elsevier Science Publishers, 1993: 33—40.

[5] OOSTHUIZEN G D. The application of concept lattice to machine learning[R]. Technical Resport University of Pretoria, South Africa, 1996.

[6] SAHAMI M. Learning classification rules using lattices[C]// Proceedings of the ECML-95. Grete: Elsevier Science Publishers, 1995: 343—346.

[7] GODIN R, MISSAOUI R. An incremental concept formation approach for learning from databases[M]. Theoretical Computer Science, 1994, 133: 387—419.

[8] FU H Y, FU H G, NJIWOUA P, et al. A Comparative Study of FCA-Based Supervised Classification Algorithms[C]// Proceedings of the Second International Conference on Formal Concept Analysis LNCS 2961. Berlin: Springer-Verlag, 2004: 313—320.

[9] CARPINETO C, ROMANO G. Galois: An order-theoretic approach to conceptual clustering[C]// Proceedings of the Tenth International Conference on Machine Learning. Amherst: Elsevier, 1993: 33—40.

[10] SAHAMI M. Learning Classification Rules Using Lattices[C]// Proceedings of the Eighth European Conference on Machine Learning. Berlin: Springer-Verlag, 1995: 343—346.

[11] MEPHU NGUIFO E. Galois Lattice: A Framework for Concept Learning Design, Evaluation and Refinement[C]// Proceedings of the Sixth International Conference on Tools with Artificial Intelligence. New Orleans: IEEE Press, 1994: 461—467.

[12] XIE Z P, LIU Z T. Research on Classifier Based on Lattice Structure [C]// Proceedings of Conference on Intelligent Information Processing. Beijing: [s.n.], 2000: 333—338.

[13] ANAMIKA G, NAVEEN K, VASUDHA B. Incremental Classification Rules Based on Association Rules Using Formal Concept Analysis [C]// Proceedings of the 4th International Conference on Machine Learning and Data Mining in Pattern Recognition. LNCS 3587. Berlin: Springer-Verlag, 2005: 11—20.

[14] GANTER B, WILLE R. Formal concept analysis[M]. Springer, 1999.

[15] 王国胤,何晓. 一种不确定性条件下的自主式知识学习模型[J]. 软件学报, 2003, 14(6): 1096—1102.

[16] MLReposi—Tory[EB/OL]. [2007—04—01]. http://www.ics.uci.edu/~mlearn/MLReposi—Tory.

[17] 侯利娟,王国胤,聂能,等. 粗糙集理论中的离散化问题[J]. 计算机科学, 2000, 27(12): 89—94.