基于类 FP-tree 的多层关联分类器

李 琳 邵峰晶 杨厚俊 孙仁诚
(青岛大学信息工程学院 青岛 266071)

摘 要 针对传统多层关联分类挖掘产生大量冗余规则而影响分类效率的问题,提出了一种基于类 FP-tree 的多层关联分类器 MACCF(Multi-level Associative Classifier based on Class FP-tree)。该分类器依据事务的类标号划分训练集,采用闭频繁模式(CLOSET+)产生完全候选项目集,通过设计适当的类内规则剪枝策略和类间规则剪枝策略,减少了大量冗余的分类规则,提高了分类的准确率;采用交叉关联规则方法,解决了交叉层数据的分类问题,实验结果表明了算法的高效性。

关键词 数据挖掘,多层关联分类器,FP-tree,剪枝,闭频繁模式
中图法分类号 TP391.4 **文献标识码** A

Multi-level Associative Classifier Based on Class FP-tree

LI Lin SHAO Feng-jing YANG Hou-jun SUN Ren-cheng
(Information and Engineering Department, Qingdao University, Qingdao 266071, China)

Abstract Focused on the problem that the traditional multi-level associative classification mining cause a lot of redundancy rules which affecte the efficiency of classification, this paper presented an improved multi-level associative classifier based on Class FP-tree named MACCF (Multi-level Associative Classifier based on Class FP-tree). It is to plot the training set based on the class property of records, using CLOSET+ generates to complete candidate item set, through the proposal inside and outside prune strategies reduce most of redundancy and has improved the accuracy; adopting cross associative method to solve the cross-level classification, experimental results show its high efficiency in classification.

Keywords Data mining, Multi-level associative classifier, FP-tree, Prune, CLOSET+

近年来,数据挖掘领域的一些学者将关联规则挖掘和数据分类相结合,提出了一种新的分类方法:关联分类[1]。关联分类通过搜索频繁模式(属性-值对的合取)与类标号之间的强关联,比诸如 C4.5 等传统的分类方法更准确。

目前,主要的关联分类算法可分为单层关联分类算法与多层关联分类算法。其中,单层关联分类算法主要有 Li Wen-min 提出的 CMAR[2], Yin Xiao-xin 提出的 CPAR[3], Cheng Hong 提出的 DDPMine[4]。CMAR 采用 FP 增长算法的变形来发现满足最小支持度和最小置信度阈值的规则的完全集。CPAR 采用基于 FOIL[5] 的方法产生规则。DDPMine 采用分支限界算法,无需产生闭频繁模式集。文献[6]支持大数据集的关联分类。

在多层关联规则挖掘的算法中,以 cumulate[7] 和 ML-T2L1[8] 最为著名。文献[9,10]也是解决同层、混合层和交叉层关联频繁模式挖掘的经典算法。在多层关联分类领域,目前的研究很少,文献[11]提出了一种基于商品分类信息的多层关联规则算法,该算法基于 Apriori 算法的改进,但它不仅会产生大量的候选项集,需要重复扫描数据库,而且对于关联分类中产生的大量冗余规则,也没有给出解决策略。针对这些问题,在综合以上算法优点的基础上,本文提出了一种基于类 FP-tree[12] 的多层关联分类器(Multi-level Associative Classifier based on Class FP-tree, MACCF),其采用闭频繁模式(CLOSET+[13])产生完全候选项目集,极大地减少了处理的数据量,同时也保证了对频繁项集挖掘的完备性。

1 相关概念与定义

在进行多层关联分类之前,需要根据原始项目表中的项目信息建立概念层次树[14],并对概念层次树进行编码,利用概念层次树的编码信息对原始数据集中的项目进行编码,生成编码事务表。为了便于描述,我们给出如下概念:

设 T 为一棵概念层次树,T_i 为 T 的第 i 层,T_{ij} 为 T 的第 i 层的第 j 个结点,其中 $i=1,2,\cdots,n, j=1,2,\cdots,m_i, n$ 为概念树的层数,m_i 为 T_i 层结点的数目。

左序编码:从 T_i 层的左端开始,依次对 T_i 中的每一组兄弟结点按连续正整数的方式进行编码。

概念层次编码树:对于 T 中的每一层,均实施左序编码后产生的树称为 T 的概念层次编码树。

概括项目编码:若项目 I 满足 T 的一条路径,则我们称该路径从根到叶所有结点编码为概括项目编码。

概括项目集:同一事务中的项目的概括项目编码的集合。

编码事务:对于事务所有项目的概括项目集。

编码事务表:采用编码事务取代原始项目表中相对应的

到稿日期:2010-09-02 返修日期:2010-12-23 本文受国家公益性行业科研专项(200905030),国家海洋局重点实验室开放基金课题(MASEG200812),山东省高等学校科技计划项目(J06G53)资助。

李 琳(1985—),男,硕士生,主要研究方向为数据挖掘和数据仓库。

事务生成的表。

为了说明编码的过程，给出如下实例。表 1 为原始项目表，表 2 是根据表 1 中的大类、成分和商标属性生成的编码事务表，T_1 为一条编码事务，T_1 中的每一项表示概括项目编码，概括项目编码是对概念层次树进行的编码。

表 1　原始项目表

项目编码	大类	商标	制造商	成分	规格	保存期	价格
17325	Milk	Foremost	Foremost Farm	2%	1 gallon	14(datys)	$3.89
…	…	…	…	…	…	…	…

表 2　编码事务表

事务	概括项目集
T1	{111,121,211,221}
…	…

例 1　对图 1 的概念层次树进行左序编码之后生成的概念层次编码树如图 2 所示。

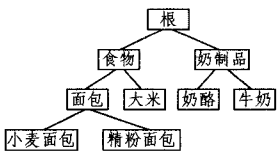

图 1　概念层次树

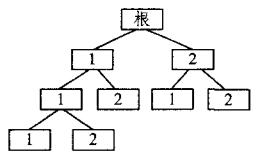

图 2　概念层次编码树

类频繁模式树：一张编码事务表建立的频繁模式树称为一棵类频繁模式树。

在同层及混合层频繁模式挖掘时，为了便于描述给出如下定义。

定义 1　如果项目 f 在编码事务表中为非 j 层频繁项目，但其父结点 e 在编码事务表中为 $j-1$ 层频繁项目，则 f 称为 e 的弥补项，记为 $Re[j,1]$，$j-1$ 层后编码变为 $*$。

定义 2　j 层 k 项目集在编码事务表中用 $A[j,k](k=1,2,\cdots,n)$ 表示。若项目 $A[j,1]$ 不是 j 层频繁项目，但 $A[j,1]$ 在跨层次挖掘时要跨到 i 层($j<i$)，在 i 层是频繁项目，则称 $A[j,1]$ 为跨 i 层频繁项目，记为 $LS[j,1,i]$，项目 $A[j,1]$ 在编码中 j 层后编码变为※。

定义 3　如果项目 f 在编码事务表中不是 j 层频繁项目和其他层的跨层次频繁项目，其父结点 e 在编码事务表中也不是 $j-1$ 层频繁项目，但 f 在编码事务表中是跨 i 层次频繁项目，则称 f 为 e 的跨层次弥补项目，记为 $LSre[j,1]$。e 编码中，$j-1$ 层后的编码变为※，i 层后编码变为☆。

2　基于类频繁模式树的多层关联分类器

2.1　数据预处理

进行多层关联分类之前，首先对训练集数据进行数据预处理，将数据转换为符合编码事务表的格式，然后针对每张编码事务表生成一棵类频繁模式树。

2.2　算法描述

使用关联规则挖掘频繁项集时，可能产生大量频繁项集，当最小支持度阈值设置较低或数据集中存在长模式时尤其如此。Han Jia-wei[13] 指出，闭频繁项集可以显著减少频繁项集挖掘所产生的模式的数量，而且保持关于频繁项集的完全信息。因此，在挖掘时挖掘闭频繁项集会减少大量数据和开销。

基于类 FP-tree 的多层关联分类器，在搜索每张编码事务表时，直接搜索闭频繁项集。在搜索的过程中采用类内剪枝策略(项合并和子项集剪枝)，为每张编码事务表生成一棵类频繁模式树。在所有的类频繁模式树都生成之后，进行类间规则剪枝，生成最终的多层关联分类规则。

在进行交叉层关联规则挖掘[14] 时，首先产生频繁 l 层 1 项目集，在生成组合频繁 l 层 k 项目集时($k>1$)，不但要使用 l 层($k-1$)项目集进行自连接，生成 l 层 k 项目候选集，还要使用 1 层 1 项目集、2 层 1 项目集、…、$l-1$ 层项目集与 l 层($k-1$)项目连接，生成 l 层 k 项目候选集。在使用各层 1 项目集与 l 层($k-1$)项目集生成 l 层 k 项目候选集时，需要将 l 层($k-1$)项目集与自身祖先连接所得到的项目集删除掉，因为 l 层($k-1$)项目集本身暗示其祖先的存在。

类内剪枝和类间剪枝策略描述如下。

2.2.1　类内剪枝

跨层次频繁模式挖掘时，采用项合并和子项集剪枝[15] 策略。项合并：如果包含频繁项目集 X 的每个事务都包含项目集 Y，但不包含 Y 的任何真超集，则 $X\cup Y$ 形成一个闭频繁项目集，并且不必再搜索包含 X 但不包含 Y 的任何项集。子项集剪枝：如果频繁项目集 X 是一个已经发现的闭频繁项目集 Y 的真子集，并且 support-count(X) = support-count(Y)，则 X 和 X 在集合枚举树中的所有后代都不可能是闭频繁项目集，因此可以剪枝。例如在表 3 中，前缀项集{323：2}的投影条件数据库是{221,211}，{221,211,111}，我们可以看出它的每个事物都包含项集{221,211}，但不包含{221,211}的真超集。项集{221,211}可以和{323}合并，形成闭项集{221,211,323：2}，并且不必挖掘包含{323}但不包含{221,211}的闭项集。

表 3　通过条件模式基产生的频繁模式树

项	条件模式基	条件 FP 树	产生的频繁模式
323	{{221 221：1}，{221 211 111：1}}	⟨221：2，211：2⟩	{221 323：2}，{211 323：2}，{221 211 323：2}
222	{{211 111：1}，{211：1}}	{211：2}	{211 222：2}

2.2.2　类间剪枝

设共有 K 个类，其中每个类是 n 条规则的析取，其中 R_k 称为类规则集。且

$$R_1=\{r_{11},r_{12},\cdots,r_{1n}\}$$

……

$$R_k=\{r_{1n},r_{2n},\cdots,r_{kn}\}$$

(1)如果类规则集之间没有交集，则不需要剪枝。

(2)类规则集之间有交集，则

①如果两个类规则集完全相等(类规则集 1 中的每一条规则与类规则集 2 中的每一条规则对应相等)，则剪枝；

策略：ⅰ)如果两个类规则集置信度不相等，则置信度高的类规则集决定样本的类别，删除置信度低的类规则集；ⅱ)如果两个类规则集置信度相等，则支持度高的类规则集决定样本的类别，删除支持度低的类规则集；

②如果一个类规则集中的一条规则是另一个规则集中一条规则的子集，则不需要剪枝；

例如：在 xx 行业会员级别中，某人符合基本的条件是普通会员，但不具备另外一个关键条件，故不能成为高级会员；

③如果恰有一条编码事务属于多个分类规则集，则采用投票的方式(概率)决定该编码事务的类别。

在类频繁模式树中，从根到叶子结点的每条路径形成一条分类规则，同一棵类频繁模式树中的多个分类规则之间是

析取的关系。

2.3 算法描述

2.3.1 MACCF 算法伪代码

输入：样本数据集 D，用户自定义各层最小支持度阈值 minsup. j ($j=1,2,\cdots,m$)

输出：多层关联分类结果集

方法：

根据类属性将样本数据集划分为 n 个编码事务表；

For($i=1;i\leqslant n;i++$){

每个类分别建立类频繁模式树；//详细描述见算法 2.3.2

输出分类规则结果集；}

End

2.3.2 类 FP-tree 生成算法伪代码

输入：n 个编码事务表

输出：类频繁模式

方法：

While(i 类中项目集不为空){

统计每个候选 1-项目集 x 计数 Sum_x^i；

If($Sum_x^i \geqslant$min count$_i(R)$)

Then {

产生 1 层频繁 1-项目集 L[1,1]及跨层次频繁项 LS[1,1,i]；

扫描 i 类编码事务表；

统计各层 1-项目集的支持数；

产生各层频繁 1-项目集及修补项，各层跨层次频繁 1-项目集及跨层次修补项；

建立项头表；

按照支持度计数降序依次将频繁项、修补项、跨层次频繁项、跨层次修补项存放在项头表中；

规则剪枝；//详细描述见 2.3.3 节

构造类 i 的 FP-tree^i

建立条件模式基，得出频繁项集；

生成基于 FP-tree^i 的多层频繁模式；//频繁模式树}}

End

2.3.3 剪枝算法伪代码

输入：未修剪项头表

输出：修剪后项头表

方法：

If(局部频繁项目 x 在不同层的多个头表中具有相同支持度)

Then 从项头表中删除项目 x；

If 频繁项目集 $X\subseteq Y$ && $support\_count(X)=support\_count(Y)$ //Y 为已发现闭频繁项目集

Then 删除项目集 X

If(类规则集 $i\cap$类规则集 $j\neq\Phi$) {

If (类规则集 $i==$类规则集 j){

If(confidence(类规则集 i)>confidence (类规则集 j)){

If(confidence (类规则集 i)== confidence (类规则集 j)){

If(support(类规则集 i)>support (类规则集 j))

Then 删除规则集 j;}

Then 删除规则集 j;}}}

If (项目集 $X\in$(类规则 i)&& 项目集 $X\in$(类规则 j)…&& 项目集 $X\in$(类规则 k))

Then 投票的方式决定项目集 X 的类别

End

3 实验分析

在 UCI 数据集上，对本文提出的 MACCF 算法进行了实验。实验证明该分类器能够实现快速的多层关联分类，与基于商品分类信息的多层关联规则挖掘算法相比，大大提高了多层关联分类挖掘的效率。

3.1 实验运行环境

硬件环境：Intel(R) Core(TM) Duo CPU 2.20GHz，3.0 GB 物理内存。

软件环境：Microsoft windows XP，SQL Server 2005 Express Edition，java。

3.2 实验结果分析

实验中截取部分数据作为分析样例，其中 1 层的最小支持数为 5，2 层的最小支持数为 3，3 层的最小支持数为 2，生成的编码事务表如表 4 所列。

表 4 编码事务表

TID	项目编码表	类别
T1	{111,121,211,221}	1
T3	{111,211,222,323}	1
T4	{112,122,221,411}	1
T6	{111,121}	1
T8	{111,122,211,221}	1
T10	{211,321,524}	1
T11	{313,411,713}	1

表 4 的编码事务表生成的类频繁模式树如图 3 所示。

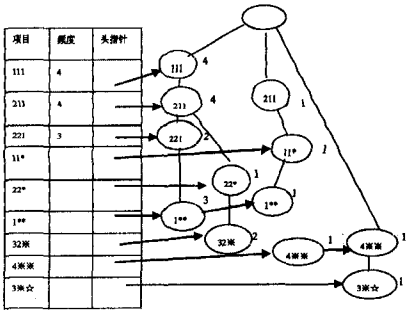

图 3 类频繁模式树

图 4 给出了各层在特定的最小支持度阈值(第 1 层支持数 5，第 2 层支持数 3，第 3 层支持数 2)的情况下随数据集大小的变化，MACCF 算法和基于商品分类信息的多层关联规则挖掘算法的执行时间比较结果。

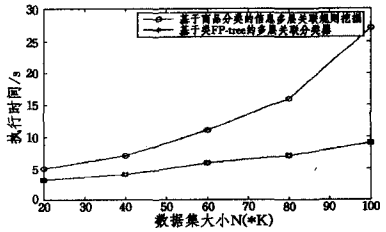

图 4 算法执行时间与数据集大小的关系

图 5 给出了在训练集大小不变的情况下，随着最小支持度的变化，MACCF 算法与基于商品分类信息的多层关联规则挖掘算法的执行时间的比较结果。

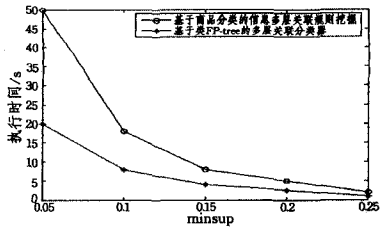

图 5 算法的执行时间与最小支持度之间的关系(第 3 层)

(下转第 211 页)

综合以上仿真实验可以看出，本文提出的算法与传统的遗传算法和量子遗传算法相比，表现出良好的快速求解性能，并且在规模较大、环境较复杂的情况下，解的质量也显示了较高的稳定性。此外，还发现只要在起始点和目标点之间有一条通道客观存在，本算法都能够规划出优化路径。

结束语 本文提出了一种基于量子遗传算法的移动机器人路径规划方法。通过仿真实验验证，该方法表现了良好的快速求解性能，并且在规模较大、环境较复杂的情况下，解的质量也显示了较高的稳定性。通过对比分析，本文算法在稳定性能、寻优能力和收敛速度方面均优于 GA 和 QGA，且对复杂寻优问题具有普遍适应性，使机器人在运动过程中避开了规划时所设置的陷阱，并且目标在障碍物附近也可以到达。

参 考 文 献

[1] 刘砚菊，杨青川，辜吟吟. 蚁群算法在机器人路径规划中的应用研究[J]. 计算机科学，2008，35(5)：263-265

[2] Stentz A C D. A real-time resolution optimal replanning for globally constraint problem [C]// The18th National Conf. on Artificial Intelligence. Cambridge. MA: MIT Press; Alberta, Canada: Edmonton, 2002: 1088-1096

[3] 刘国栋，谢宏斌，李春光. 动态环境中基于遗传算法的移动机器人路径规划的方法[J]. 机器人，2003，25(4)：327-330

[4] Khatib O. Real-time obstacle avoidance for manipulators and mobile robots [J]. Int J of Robotic Research, 1986, 5(1): 90-98

[5] 王醒厕，张汝波，顾国昌. 基于势场栅格法的机器人全局路径规划[J]. 哈尔滨工程大学学报，2003，24(2)：170-173

[6] Narayanan A, Moore M. Quantum-inspired genetic algorithm [C]// Proc of IEEE International Conference on Evolutionary Computation. Piscataway: IEEE Press, 1996: 61-66

[7] Han K-H, Kim J-H. Genetic quantum algorithm andits application to combinatorial optimization problem[C]// Proceedings of the 2000 Congress on Evolutionary Computation. 2000: 1354-1360

[8] Koren Y, Borenstein J. Potential field methods and their inherent limitations for mobile robot navigation [C]// Proc. IEEE Conf. Robotics and Automation. Sacramento, CA, Apr. 1991: 1398-1404

[9] Han K-H, ParkK-H, Lee C-H, et al. Parallel quantum-inspired genetic algorithm for combinatorial optimization prob-lem[A]// Proceedings of the 2001 Congress on Evolutionary Compu-tation [C]. USA: IEEE Press, 2001: 1422-1429

[10] 杨俊安，庄镇泉，史亮. 多宇宙并行量子遗传算法[J]. 电子学报，2006，32(6)：923-928

(上接第 178 页)

从图 4 可以看出，随着数据集的增加，MACCF 算法的执行时间基本是线性变化的，算法的可伸缩性较好。从图 5 的分析可知，在数据集大小不变的情况下，与基于商品分类信息的多层关联规则挖掘算法相比，MACCF 算法大大降低了算法的执行时间。

结束语 本文针对基于商品分类信息的多层关联规则算法不但会产生大量的候选项集，需要重复扫描数据库，而且对于关联分类中产生的大量冗余规则，没有给出解决策略的问题，提出了基于类 FP-tree 的多层关联分类器 MACCF 算法，并提出了类内、类间规则剪枝策略，从而大大提高了多层关联分类挖掘算法的执行效率。

参 考 文 献

[1] Liu Bing, Hsu W, Ma Yi-ming. Integrating Classification and Association Rule Mining[C]// Proceedings of KDD. 1998: 80-86

[2] Li W, Han J, Pei J. CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules[C]// Proc. 2001 Int. Conf. on Data Mining (ICDM'01). San Jose, CA, Nov. 2001

[3] Yin X, Han J. CPAR: Classification based on Predictive Association Rules[C]// Proc. 2003 SIAM Int. Conf. on Data Mining (SDM'03). San Fransisco, CA, May 2003

[4] Cheng Hong, Yan Xi-feng, Han Jia-wei, et al. Direct Discriminative Pattern Mining for Effective Classification[C]// Proc. 2008 Int. Conf. on Data Engineering (ICDE'08). Cancun, Mexico, April 2008

[5] Quinlan J R, Cameron-Jones R M. FOIL: Midterm Report[C]. 2006

[6] Cheng Hong, Yan Xi-feng, Han Jia-wei, et al. Discriminative Frequent Pattern Analysis for Effective Classification[C]// Proceedings of the 2007 IEEE International Conference on Data Engineering (ICDE 07). Istanbul, Turkey, April 2007

[7] Srikant R, Agrawal R. Mining Generalized Association Rules[C]// 21st Int'l Conference on Very Large Databases. Sep. 1995

[8] Han J, Fu Y. Discovery of Multi-level Association Rules from Large Databases [C]// Proceedings of the 21st International Conference on VLDB. Zurich, Switzerland, 1995

[9] Thakur R S, Jain R C, Pardasani K R. Fast Algorithm for Mining Multi-Level Association Rules in Large Databases [J]. Asian Journal of Information Management, 2007, 1(1): 19-26

[10] Thakur R S, Jain R C, Pardasani K R. Mining Level-crossing Association Rules from Large Databases [J]. Journal of Computer Science, 2006, 2(1): 76-81,

[11] 鲁增秋，陈玉哲，王殿升. 一种改进的基于商品分类信息的多层关联规则挖掘算法[J]. 科技情报开发与经济，2006，16(14)：137-139

[12] Han J, Pei J. Mining Frequent Patterns by Pattern-growth: Methodology and Implications [C]// ACM SIGKDD Explorations (Special Issue on Scaleble Data Mining Algorithms). December 2000, 2(2)

[13] Wang J, Han J, Pei J. CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets[C]// Proc. 2003 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03). Washington D C, Aug. 2003

[14] 邵峰晶，于忠清，王金龙，等. 数据挖掘原理与算法(第二版)[M]. 北京：科学出版社，2009

[15] 韩家炜. 数据挖掘概念与技术(第二版)[M]. 北京：机械工业出版社，2008