计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：　决策树剪枝加强的关联规则分类方法
作者：　范劭博，张中杰，黄健
收稿日期：　2022-06-30
网络首发日期：　2022-09-17
引用格式：　范劭博，张中杰，黄健．决策树剪枝加强的关联规则分类方法[J/OL]．计算机工程与应用. https://kns.cnki.net/kcms/detail/11.2127.TP.20220916.1612.014.html

网络首发时间：2022-09-17 14:29:12
网络首发地址：https://kns.cnki.net/kcms/detail/11.2127.TP.20220916.1612.014.html

决策树剪枝加强的关联规则分类方法

范劭博，张中杰，黄健

国防科技大学 智能科学学院，长沙 410073

摘 要：传统关联规则挖掘在面临分类决策问题时，易出现非频繁规则遗漏，预测精度不高的问题。为得到正确合理且更为完整的规则，提出了一种改进方法 DT-AR(Decision Tree-Association Rule Algorithm)，利用决策树剪枝策略对关联规则集进行补充。首先，该方法利用 FP-Growth(Frequent Pattern Growth)算法得到关联规则集，其次利用 C4.5 算法构建后剪枝决策树并提取分类规则，在进行置信度迭代筛选后与关联规则集取并集修正，最后利用置信度作为权重系数采取投票法进行分类。实验表明，与传统关联规则挖掘和决策树剪枝方法相比，该方法得到的规则在数据集分类结果上更准确。

关键词：数据挖掘；决策树剪枝；关联规则分类；数据分类
文献标志码: A **中图分类号**:TP391 doi：10.3778/j.issn.1002-8331.2206-0476

The Association Rule Classification Method Strengthened by Decision Tree Pruning

FAN Shaobo, ZHANG Zhongjie, HUANG Jian

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China

Abstract：Solving the problem of mining rules assisting decision making with traditional association rule comes along with omission of infrequent rules and the low prediction accuracy usually. In order to obtain more correct, reasonable and complete rules, a method, DT-AR(Decision Tree-Association Rule Algorithm) is proposed, which supplements the association rule set by decision tree pruning strategy. Firstly, the method gets the association rule set by FP-Growth(Frequent Pattern Growth).And then it constructs the post pruning decision tree and extracts classification rules. After iterative confidence filtering, it makes corrections to the association rule set to take the union with the classification rule set. Finally, it makes data classification by voting method, with the conifidence as the weight. The experiment results show that the rules of DT-AR get more accurate results in the prediction compared with the method of association rule and decision tree pruning.

Key words：data mining; decision tree pruning; association rule classification; data classification

关联规则挖掘是数据挖掘的重要技术，常被用以挖掘数据集中的共现规律，从而进行辅助决策[1]。然而，现有的关联规则技术在进行辅助决策时，存在过拟合以及非频繁规则的遗漏问题，导致关联规则提取不完整，预测精度降低。

针对非频繁且有意义的规则，可通过决策树学习方法进行补充。决策树是一种常见的机器学习分类方法，用以从给定的数据集学得一个模型，从而对新示

基金项目：国家自然科学基金(“多采样策略下的强精度近似频繁模式挖掘”(61906202))。
作者简介：范劭博(1998－),男,硕士生,研究方向为数据挖掘,E-mail:2549633209@qq.com;张中杰(1988－),男,博士,讲师，研究方向为数据挖掘;黄健(1971－)，女，博士生导师，研究员，研究方向为任务规划、分布式仿真。
收稿日期：2022-06-30；

例分类，基本流程遵循“分而治之”的递归处理机制，具有较强的泛化能力，能够较好的处理未见示例[2]。

采取剪枝策略可以在一定程度上改善过拟合，故针对过拟合问题，可借鉴剪枝策略思想，对关联规则和决策树分类规则在训练集上再进行一轮迭代置信度筛选，进一步改善分类规则的过拟合。

综上，本文提出一种算法决策树剪枝加强的关联规则分类算法 DT-AR，针对规则过拟合和遗漏问题对 FP-Growth 算法进行改进，利用决策树剪枝策略提取分类规则，并对关联规则集进行修正补充，旨在提取更完整的关联规则集，提高规则分类精度。

1 相关研究

目前关联规则挖掘和决策树都是数据分类的经典方法，能够取得不错的分类效果，但是存在着明显的缺点：关联规则算法对频繁项的搜索能够有效获取规则，但同时也易出现规则遗漏；决策树方法利用剪枝提高了分类精度，但是规则数量明显减少。且两者都存在着规则过拟合的问题。近年来，相关研究针对算法分类效果，从不同的方面入手，对上述方法进行研究和改进，优化了算法流程，在一定程度上提高了分类精度。

关联规则挖掘是数据挖掘技术的重要研究内容，其核心任务为通过递归搜索数据库，得到满足最小支持度的项集。目前，关联规则挖掘已被运用于数据分类中，基本原理是将属性和类别标签联系在一起，强调频繁模式与类别标签之间的关联。文献[3]最早提出基于关联规则的分类方法 CBA(Classification Base of Association)，将关联规则挖掘和分类规则挖掘方法结合，提出一种基于已发现类关联规则集合构建分类器的有效算法。

目前关联规则分类的大部分研究仍然针对提升方法的分类精度或者算法效率，如文献[4]提出了一种基于新剪枝度量的关联分类器，规则精度索引分类器，不同于CBA使用置信度作为度量，该方法使用规则精度指数作为度量对规则进行修剪，提高了规则修剪效率和分类准确率，但对规则过拟合问题考虑欠缺。文献[5, 6]同样针对分类器的效率和精度开展了研究，提高分类器的整体覆盖率，在分类准确率上表现较好，在一定程度上改善了过拟合的问题。而一些研究从规则生成方法和数据结构方面入手进行优化，在提高分类精度的同时能够改善过拟合问题，如文献[7]为解决CBA算法的基础上挖掘定量关联规则，提出了一种基于定量关联规则树的分类及回归预测算法，提高了分类准确率并降低了计算复杂度。文献[8]提出了一种新的不均衡关联分类算法 ACI(Associative Classification Algorithm for Imbalanced Data)，利用不均衡规则裁剪方法对规则进行裁剪，并存储到CR树中用于分类，通过改善规则生成流程提升了分类精度。文献[9]提出了一种改进的带约束的关联分类规则挖掘方法，采用扩展概念格结构存储频繁项集，引入差集概念提升计算速度，依据给定约束条件提取关联分类规则，明显提高了算法效率和分类效果。上述方法在算法的分类效率和精度上均有了明显提升，并在一定程度上改善了规则的过拟合问题，但是未过多考虑关联规则的遗漏问题。

为减少规则的遗漏，相关研究从稀有项以及稀有规则入手，如文献[10]提出了一种改进的分类关联规则挖掘方法，为每个规则设定最小支持度，既有效避免了产生过多冗余规则，又能对某些低频规则顺利挖掘。文献[11, 12]针对提取高置信度的稀有关联规则开展了研究，能够有效挖掘传统关联规则方法遗漏的重要规则，使得关联规则集更加完整。

决策树分类算法是利用数据集生成一棵在特征空间上的条件概率分布树，是一系列关联规则的集合，侧重于数据类别之间的差异，可用于对数据集分类。在生成决策树时，决策树分支节点所包含的样本应尽可能属于同一类别，通常需要一定指标来表示特征的分类能力，信息熵就是最常用的一种指标，用来表示随机变量不确定性的度量。典型决策树分类算法包括 ID3(Iterative Dichotomiser 3)、C4.5(Iterative Dichotomiser 4.5)、CART(Classification And Regression Tree)等，分别采取不同的标准划分属性，其中 ID3 使用信息增益对数据集进行划分，C4.5 在前者基础上采取增益率，而 CART 则采取基尼系数作为度量标准，本文采取 C4.5 算法，在构建决策树时计算增益率划分属性。

为了提高决策树面对未知数据的泛化能力，通常需要对决策树进行剪枝操作，决策树的剪枝策略可分为预剪枝和后剪枝，预剪枝在生成决策树时预先对节点进行估计，而后剪枝则是生成决策树后自下至上进行对非叶节点进行考察和剪枝操作。其中后剪枝策略可以有效避免欠拟合问题的发生。

决策树剪枝分类的相关研究同样面向提高分类精度和决策树构建效率，如文献[13]提出了一种基于分数近似算法的决策树分类算法 DTFA(Decision Trees based on the Fractions Approximation Algorithm)，采取

全新的分类准则，使用霍夫丁不等式获取所需边界，提高了分类精度，但过拟合问题仍然较明显。此外，一些研究针对决策树的构建方法进行分析，通过建立更有效的决策树结构提高分类效果，如文献[14]提出一种基于代价敏感集成决策树的分类方法，通过构建继承决策树生成更准确的分类器，引入代价敏感因子为不同分类结果赋予权重，提升了非频繁类的分类识别率。文献[15]提出了一种基于 bagging(Bootstrap Aggregating)思想的集成决策树分类算法，随机采样抽取样本，选取最佳分类属性作为节点，直至决策树构建完毕，该方法在分类准确率上有明显提升，并且提高了算法效率。与上述文献专注提高算法分类精度不同，文献[16]针对本文提出的过拟合与规则遗漏问题，提出了一种改进多决策树方法，利用欠采样技术处理数据集生成平衡子集，在每个子集上构建单决策树最后集成多决策树，在一定程度上避免了过拟合和信息丢失的问题。

目前关联规则挖掘算法的相关研究主要面向提高算法的效率，分类效果虽然有了明显提升，但是对规则过拟合和遗漏问题的考虑仍然不足。决策树分类方法的相关研究主要面向决策树的构建方法，虽然部分算法能够提高非频繁数据类的分类精度，但是过拟合问题仍明显存在。本文面向关联规则分类决策问题，针对规则过拟合和遗漏问题，在关联规则挖掘 FP-Growth 算法的基础上，利用决策树剪枝策略提取规则，采用一种新的集成方式对关联规则进行修正补充，并利用贝叶斯法则计算置信度作为规则权重系数，实现了决策树剪枝加强的关联规则提取方法 DT-AR，提取了更完整的关联规则集，提高了规则的分类精度。

2 DT-AR 算法

2.1 问题描述

设事务数据 $DB=\{D_1、D_2、\ldots、D_l\}$为数据集，其包含若干属性类别和标签类别。每一条事务数据 D_x 由若干属性值和一个标签值构成，$DA=\{A_1、A_2、\ldots、A_m\}$为数据属性集合，$DL=\{L_1、L_2、\ldots、L_n\}$为数据标签集合，$DT=\{T_1、T_2、\ldots、T_m\}$为项的集合，其中 T_x 即为属性 A_x 所有取值的集合。利用关联规则挖掘解决辅助决策问题，核心任务为找到各类标签对应的频繁模式集，形成关联规则，并利用其实现对数据的预测分类。

2.2 相关定义

首先对相关参数定义如下：

m：数据属性数目；

n：数据分类标签数目；

$L_1、L_2、\ldots、L_n$：数据分类标签，共分为 n 类；

$A\{A_1、A_2、\ldots、A_m\}$：数据属性集合(或事件集合)，共 m 个种类；

D：总体初始数据集；

$D_S=D_S^1\cup D_S^2\cup\ldots\cup D_S^n$：数据样本集总集，用于构建初始决策树，其中 $D_S^i(i=1、2、\ldots、n)$表示第 i 个标签对应的子数据样本集；

$D_V=D_V^1\cup D_V^2\cup\ldots\cup D_V^n$：数据验证集总集，用于决策树剪枝时计算精度，其中 $D_V^i(i=1、2、\ldots、n)$表示第 i 个标签对应的子数据验证集；

$D_E=D_E^1\cup D_E^2\cup\ldots\cup D_E^n$：数据训练集总集，用于关联规则挖掘，且 $D_E=D_S\cup D_V$，其中 $D_E^i(i=1、2、\ldots、n)$表示第 i 个标签对应的子数据训练集；

$D_P=D_P^1\cup D_P^2\cup\ldots\cup D_P^n$：数据预测集总集，用于实验验证分析，其中 $D_P^i(i=1、2、\ldots、n)$表示第 i 个标签对应的子数据预测集；

minsup：最小支持度阈值；

minconf：最小置信度阈值；

sup(X)：项或模式 X 的支持度，为数据库中包含 X 的数目占总体数据数目百分比；

conf(X,L_X)：规则“$X\rightarrow L_X$”的置信度，为数据库中标签 L_X 关于模式 X 的后验概率；

$S_{DT}=S_{DT}^1\cup S_{DT}^2\cup\ldots\cup S_{DT}^n$：采用关联规则挖掘方法得到的模式集,且 $S_{DT}^i(i=1、2、\ldots、n)$表示从第 i 个标签对应数据中挖掘而来的频繁模式集；

$S_{PDT}=S_{PDT}^1\cup S_{PDT}^2\cup\ldots\cup S_{PDT}^n$：采用决策树剪枝方法得到的分类规则集,且 $S_{PDT}^i(i=1、2、\ldots、n)$表示分类为第 i 个标签的规则子集；

$S_{FR}=S_{FR}^1\cup S_{FR}^2\cup\ldots\cup S_{FR}^n$：修正后的最终规则集,且 $S_{FR}^i(i=1、2、\ldots、n)$表示第 i 个标签对应的子规则集；

l-模式：长度为 l 的模式。

另外对相应概念定义如下：

定义 1 决策树分类规则：决策树每一条完整分支的从上至下节点组合。

定义 2 频繁模式：支持度满足最小支持度阈值要求 $sup(X)>minsup$ 的项集。

定义 3 数据分类规则：设第 x 个标签为 L_x，对于 S_{DT}^x、S_{PDT}^x 或 S_{FR}^x 中的模式 X，其对应分类规则为“$X\rightarrow L_x$”，表示“具有模式 X 的样本，其数据标签应为 L_x”。

2.3 DT-AR 算法

本文提出了一种决策树剪枝加强的关联规则分类方法，首先利用挖掘算法处理数据集得到频繁模式，

经置信度筛选后生成关联规则集，再通过构建后剪枝决策树得到决策树分类规则集，与关联规则集取并集，进行置信度迭代筛选得到最终规则集。在数据预测时，采用置信度作为规则加权系数进行分类。算法的流程示意如图 1 所示。

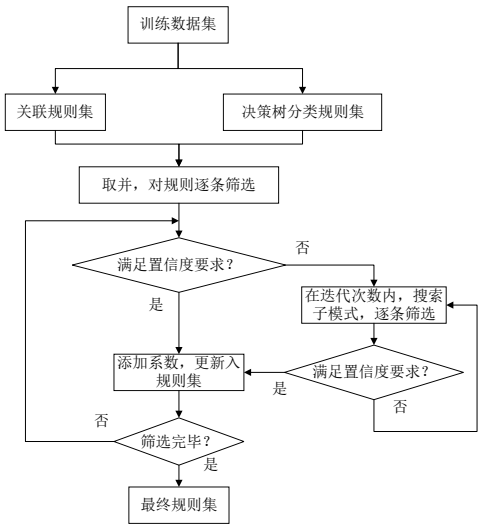

图 1 DT-AR 算法流程示意图

Fig.1 The schematic diagram of DT-AR algorithm

在对规则集进行置信度筛选时，本文借鉴减错剪枝策略对规则的处理思想，对每一条规则进行置信度迭代筛选，流程如下：

流程 1：置信度迭代筛选 conf_scr(*a*)

输入：规则 *a*、最大迭代次数 *k*、最小迭代长度 *len*

输出：规则集合 *A*

步骤：

判断 *a* 置信度是否满足要求

若 *conf*(*a*)>*minconf*

将 *a* 更新入 *A*

否则，判断 *a* 的长度 *l*，若 *l*>*len*

若迭代次数满足最大迭代要求

搜索 *a* 的(*l*-1)-模式集合 *B*，迭代次数自增 1

若 *B*≠∅

对 *B* 中的每一条规则 *b*，执行 conf_scr(*b*)

返回 *A*

置信度计算采用公式(1)，表示对应标签结果的后验概率：

$$P(L_i \mid X)=\frac{P(X \mid L_i)\cdot P(L_i)}{\sum_{j=1}^{n} P(X \mid L_j)\cdot P(L_j)} \tag{1}$$

改进算法 DT-AR 具体步骤如下：

流程 2：算法 DT-AR

输入：数据集 *D*

输出：关联规则集 S_{FR}

步骤：

1.数据预处理，将 *D* 随机划分生成 D_S、D_V、D_P，提取标签集 $L(L_1$、L_2、…、$L_n)$并分类；

2.扫描数据集 $D_E=D_S\cup D_V$，关联规则挖掘，获取满足 *minsup* 要求的频繁模式集合：

$$S_{DT}{}' = S_{DT}{}^{1'} \cup S_{DT}{}^{2'} \cup \cdots \cup S_{DT}{}^{n'};$$

3.对 S_{DT}' 进行置信度筛选，得到规则集合：

$$S_{DT}{}^{i} = \left\{ \{x\} \mid x \in S_{DT}{}^{i'}, conf(x) > minconf \right\};$$

4.扫描数据集 D_S，计算信息增益构建决策树，用 D_V 作验证集进行剪枝，按定义 1 提取得到分类规则集：

$$S_{PDT}{}' = S_{PDT}{}^{1'} \cup S_{PDT}{}^{2'} \cup \cdots \cup S_{PDT}{}^{n'};$$

5.对 S_{PDT}' 中每条规则 *y*，执行 conf_scr(*y*)，得到 $S_{PDT}\{S_{PDT}{}^1$、$S_{PDT}{}^2$、…、$S_{PDT}{}^n\}$；

6.对每类标签 L_i，$S_{FR}{}^i=S_{DT}{}^i\cup S_{PDT}{}^i$，以规则置信度作为加权系数，生成最终规则集 S_{FR}。

决策树利用信息增益划分分类属性，而关联规则挖掘利用数据支持度进行频繁模式搜索并筛选。二者原理不同，规则特点也不同，决策树分类规则具备较高的信息增益，而关联规则数据频繁程度较高，故决策树分类规则集必然包含关联规则集遗漏的部分非频繁规则，而后者也必然包含前者因剪枝遗漏的规则，故对决策树分类规则集和关联规则集取并集可得到更完整的规则集。

由 DT-AR 方法得到关联规则集 $S_{FR}=S_{DT}\cup S_{PDT}$，故有

$$S_{DT} \subseteq S_{FR} \tag{2}$$

且 S_{PDT} 中的规则经过置信度筛选，故 DT-AR 方法能够得到更完整的规则集。

在关联规则分类中，由于过度重视频繁项，规则过拟合通常表现为规则包含更多高频项，而部分数据在分类时无法精准匹配规则，易出现错误分类或者无法分类的情况，而 DT-AR 算法在关联规则得基础上，利用决策树分类规则对规则集进行补充，完善规则集得同时在一定程度上缓解了规则过拟合的问题。

在决策树剪枝过程中，通常以验证集精度作为依据，本文在其基础上，对规则进行置信度迭代筛选，并利用整体训练集的精度作为依据，进一步降低了过拟合的风险，得到的规则具备更高的泛化能力。综上，针对规则过拟合和规则遗漏的问题，DT-AR 算法从理论上是有一定解决效果的。

2.4 时空复杂度分析

DT-AR 算法侧重于提高分类精度和规则完整性，为衡量 DT-AR 算法的运行效率和消耗代价，需对其进行时空复杂度分析。

针对时间复杂度，采取大 O 表示法分析其时间复

杂度。DT-AR 算法的核心步骤为规则的置信度迭代筛选，由算法流程 1 可知其存在一层循环嵌套，故时间复杂度为 $O(t^2)$，其中 t 为决策树分类规则的数目，而 FP-Growth 算法时间复杂度为 $O(n^2)$，其中 n 为数据集事务数，C4.5 决策树的时间复杂度为 $O(l*\log(l)*d)$，其中 l 为训练样本数，d 为数据维数。综上，DT-AR 算法的时间复杂度为 $O(n^2)$，说明 DT-AR 与 FP-Growth 算法的时间复杂度接近。

针对空间复杂度，FP-Growth 算法的空间开销主要是搜索频繁模式，其空间复杂度取决于为头表的元素搜索路径时的空间消耗，而 C4.5 决策树的空间复杂度仅取决于树的最大深度，为 $O(n)$，因此 DT-AR 算法的空间复杂度为 $O(n^2)$，与 FP-Growth 算法相同。

通过以上分析可知，DT-AR 算法的时空复杂度均与 FP-Growth 保持一致，说明 DT-AR 算法能够在保持时空复杂度不变的条件下提高分类精度。

2.5 样例说明

以表 1 中的数据集为例，首先，按标签“Y”、“N”将其分为两个数据集，并分别进行关联规则挖掘，设定最小支持度阈值为 3，得到频繁模式及对应支持度，如表 2 所示。

表 1　示例数据集

Table 1　The sample data set

编号	A	B	C	D	E	F	L
1	A1	B1	C1	D1	E1	F1	Y
2	A2	B1	C2	D1	E1	F1	Y
3	A1	B1	C1	D3	E3	F2	N
4	A2	B1	C1	D2	E1	F1	N
5	A2	B2	C2	D2	E1	F1	N
6	A2	B2	C2	D1	E2	F2	Y
7	A2	B1	C1	D1	E1	F1	Y
8	A1	B1	C2	D1	E1	F2	Y
9	A1	B2	C2	D2	E2	F1	N
10	A2	B1	C2	D1	E3	F2	N

表 2　基于表 1 生成的数据挖掘结果

Table 2　The data mining results generated based on Table1

标签	频繁模式	支持度
Y	D1	5
	B1、E1、D1-B1、D1-E1、B1-E1、D1-B1-E1、F1、A2、C2、D1-F1、B1-F1、D1-B1-F1、E1-F1、D1-E1-F1、B1-E1-F1、D1-B1-E1-F1、D1-A2、D1-C2	4
		3
N	B1、A2、D2、C2、F1、D2-F1	3

从挖掘结果可发现，未经置信度筛选的模式特征较为杂乱，“Y”、“N”两个标签对应的频繁模式集中均包括有“B1”、“A2”、“C2”、“F1”。现利用公式(1)计算规则置信度，最小置信度阈值设为 70%，经过筛选初步得到关联规则并写入关联规则集 $S_{DT}{}^1$、$S_{DT}{}^2$ 中，结果如表 3 所示。

表 3　经置信度筛选的关联规则

Table 3　The association rules filtered by confidence

标签	关联规则	置信度
Y	D1-E1→Y、B1-E1→Y、D1-B1-E1→Y、D1-F1→Y、D1-B1-F1→Y、D1-B1-E1-F1→Y	1
	D1-B1→Y	4/5
	B1-F1→Y、B1-E1-F1→Y、D1-A2→Y、D1-C2→Y	3/4
	D1→Y	5/7
N	D2→N、D2-F1→N	1

得到关联规则后，利用数据集构建决策树，提取决策树分类规则，现将数据集随机划分为样本集和验证集：令前 50%为样本集，其余为验证集。利用后剪枝策略构建分类决策树，如图 2 所示。

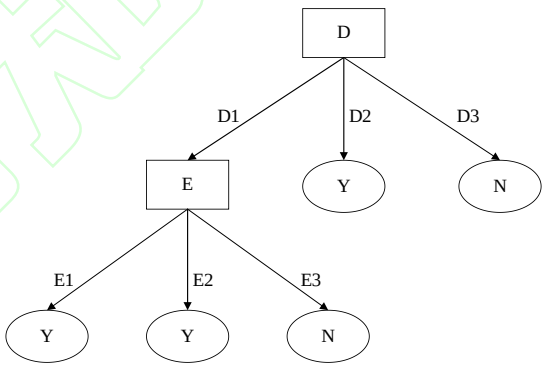

图 2 基于表 1 生成的后剪枝决策树

Fig.2 The post-pruning decision tree generated based on Table1

由决策树得到的数据分类规则集分别为 $S_{PDT}{}^1$：{“D1-E1”、“D1-E2”}、$S_{PDT}{}^2$：{“D1-E3”、“D2”、“D3”}，经计算均满足最小置信度阈值要求，与关联规则集 $S_{DT}{}^1$、$S_{DT}{}^2$ 按标签分别取并集，得到最终规则集 S_{FR}，并添加权重系数。

表 4　示例预测数据集

Table 4　The sample prediction data set

编号	A	B	C	D	E	F	L
1	A1	B2	C1	D1	E3	F2	Y
2	A2	B1	C2	D1	E3	F1	Y
3	A1	B2	C1	D1	E2	F2	N
4	A2	B2	C1	D3	E1	F1	N
5	A2	B2	C2	D3	E1	F1	N
6	A2	B2	C2	D1	E1	F2	Y
7	A2	B1	C1	D1	E1	F1	Y
8	A1	B1	C2	D1	E1	F2	Y
9	A1	B2	C2	D2	E2	F1	N
10	A2	B1	C2	D1	E3	F2	N

分别利用 FP-Growth、C4.5 决策树剪枝和 DT-AR 方法得到的规则集对表 4 中的预测数据进行分类，将结果对比如表 5 所示。

表 5　示例预测数据集分类结果

Table 5　The classification result of the sample prediction data set

方法	FP-Growth	C4.5 决策树	DT-AR
正确	6	7	8
错误	1	2	1
无法预测	3	1	1

从结果中可以发现改进算法得到的分类结果更准确，经分析后可知，FP-Growth 得到的关联规则集预测时，第 4、5 条数据未覆盖任何规则，而改进算法规则集更为完整，故对应结果分类正确。

DT-AR 方法应用到实际问题时，可能存在满足多种规则(即面临规则冲突)而难以得到确定分类结果的问题，DT-AR 方法通过计算权重系数和来确定最终分类结果，其具体流程如图 3 所示。

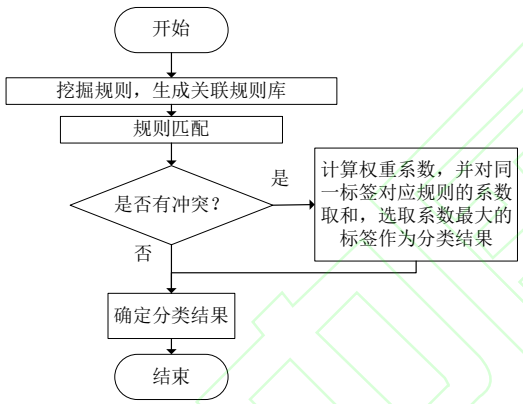

图 3　规则冲突问题处理分类流程

Fig.3　The classification process of the rule conflict problem

以示例预测数据集为例，如表 4 中第 5 条数据匹配的规则包括：1)“A2→Y”、2)“D3→N”、3)“F1→Y”，在分类时面临规则冲突问题。利用式(1)计算其权重系数分别为 75%、100%、100%，规则 1)和 3)为同一标签对应得规则，对其权重系数求和得 Q_Y=175%，而 Q_N=100%，$Q_Y > Q_N$，故该数据的分类结果为“Y”。

3 实验验证分析

3.1 实验数据参数

数据来源：实验选取互联网数据平台上公开的数据集：

1.IBM 公开的综合数据集 IBM HR Analytics Employee Attrition & Performance，下文简称 HR。

2.UCI 数据库中，来自南斯拉夫卢布尔雅那大学医疗中心肿瘤研究所的乳腺癌数据集 Breast Cancer Wisconsin Data Set，下文简称 Breast。

3.TCIA 提供的淋巴结公开数据集 CT Lymph Nodes，下文简称 Lymph。

4.UCI 数据库中的动物园数据集 Zoo Data Set，下文简称 Zoo。

5.UCI 数据库中的红酒数据集 Wine Data Set，下文简称 Wine。

实验验证选取分类标签以及数据属性进行挖掘分析，必要时对连续值属性的数据按值进行分类。实验将数据集随机划分为训练集和预测集，同时面向决策树剪枝进一步将训练集拆分为样本集和验证集。数据参数情况如表 6 所示。

表 6　实验数据参数

Table 6　The parameter of experimental data

数据集	数目	数据类别数目	训练集数目	样本集数目	验证集数目	预测集数目
HR	400	2	250	150	100	150
Breast	170	2	120	70	50	50
Lymph	142	2	92	60	32	50
Zoo	101	7	71	41	30	30
Wine	178	3	130	90	40	48

3.2 实验结果

实验对生成的关联规则进行检验分析，采用投票法决定分类结果，对每个样例计算覆盖规则数并依此分类，规则数目最多的标签记为该样例的预测结果。若某样例没有任何一条规则可以覆盖，则该样例分类结果为“无法预测”。

为了验证算法的有效性，实验采取六种分类算法作为对比，分别是 C4.5 决策树算法、FP-Growth 算法、CBA 算法、CMAR 算法、SVM(支持向量机，Support Vector Machine)算法以及朴素贝叶斯分类方法，针对每个数据集，运行 FP-Growth 算法、决策树分类算法、CBA 算法、CMAR 算法、SVM 算法和本文提出的 DT-AR 算法得到规则集，利用投票法分析预测集的标签，然后应用朴素贝叶斯方法进行分类预测，所有方法均采用留出法进行预测验证。

首先，DT-AR、CBA、CMAR、FP-Growth 算法在挖掘时均需要设定最小支持度阈值和最小置信度阈值，为得到最佳实验效果，需要进行对比实验，运用控制变量法获取最佳最小支持度阈值和最小置信度阈值。对于最小支持度阈值的确定，不妨先将最小置信度阈值设定为 50%，选取 Zoo 数据集，分别利用上述算法，设定不同最小支持度阈值进行规则挖掘及分类，其结果分别如图 4 所示。

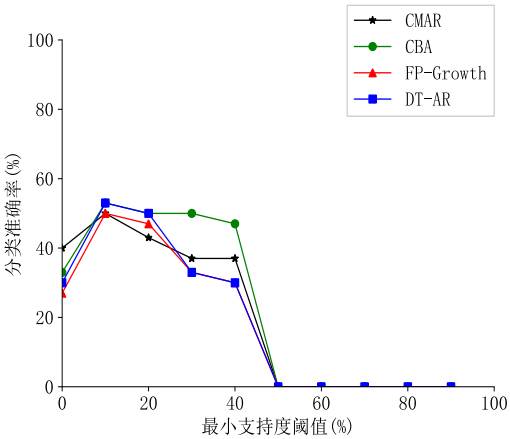

图 4 minconf=50%时各算法分类结果对比图

Fig.4 Thecomparison of classification result of the algorithms when minconf=50%

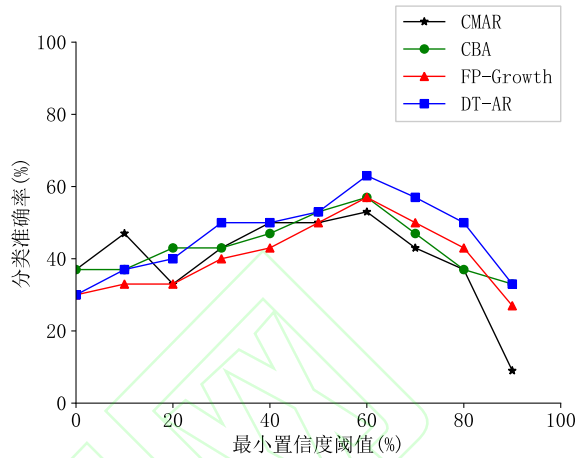

图 5 minsup=10%时各算法分类结果对比图

Fig5. The comparison of classification result of the algorithms when minsup=50%

可以发现最小支持度阈值对分类结果影响较大，取值为 10%时算法取得相对最佳效果，这是由于最小支持度阈值取值过高会导致过多信息丢失，导致分类效果不理想，过低将导致规则冗余，分类时无法准确匹配到有效规则。对于最小置信度阈值，将最小支持度阈值设定为 10%，选取 Zoo 数据集，分别利用上述算法，设定不同最小置信度阈值进行规则挖掘及分类。其结果如图 5 所示。

可以发现最小置信度阈值为 60%时算法取得最佳效果，通过分析可知，最小置信度阈值设定过高会加剧规则过拟合的问题，过低则出现规则冗余及关联性降低的问题。

综上，实验在最小支持度阈值为 10%，最小置信度为 60%的条件下进行能够取得较好的效果，将所有算法的分类结果一并汇总，如表 7 所示。

表 7 实验预测结果对比

Table 7 The comparison of experimental prediction results

方法	指标	FP-Growth	C4.5 决策树	CBA	CMAR	SVM	DT-AR	朴素贝叶斯
HR	生成规则数	131	21	123	144	/	147	/
	正确	88	80	82	88	90	93	87
	错误	45	48	68	62	60	50	63
	无法预测	17	22	0	0	0	7	0
Breast	生成规则数	55	10	56	59	/	63	/
	正确	31	32	32	34	25	35	34
	错误	16	12	18	16	25	15	16
	无法预测	3	6	0	0	0	0	0
Lymph	生成规则数	41	11	46	30	/	47	/
	正确	30	24	31	32	30	33	32
	错误	16	20	19	18	20	17	18
	无法预测	4	6	0	0	0	0	0
Zoo	生成规则数	286	6	42	301	/	289	/
	正确	17	18	17	16	19	19	18
	错误	13	12	13	14	11	11	12
	无法预测	0	0	0	0	0	0	0
Wine	生成规则数	71	17	43	76	/	87	/
	正确	37	36	35	36	19	39	38
	错误	11	12	13	12	29	9	10
	无法预测	0	0	0	0	0	0	0

3.3 实验结果分析

结合表 5 中的数据，将各算法的分类准确率对比汇总，准确率计算见公式(3)，结果如图 6 所示。通过分析，相比于其他算法，DT-AR 方法在预测准确率上具有明显的提升，且从生成规则数量方面可以发现，利用 DT-AR 方法可以得到更完整的数据分类规则集。

$$P=\frac{n(\text{预测正确数目})}{n(\text{预测样本数目})}\times 100\% \tag{3}$$

通过分析实验结果，相比于 FP-Growth 算法，采取 DT-AR 方法使得规则集覆盖样例更多,减少了无法预测的样例数目。这说明改进方法得到的关联规则集更为完整，在一定程度上避免了非频繁但重要规则的遗漏问题。

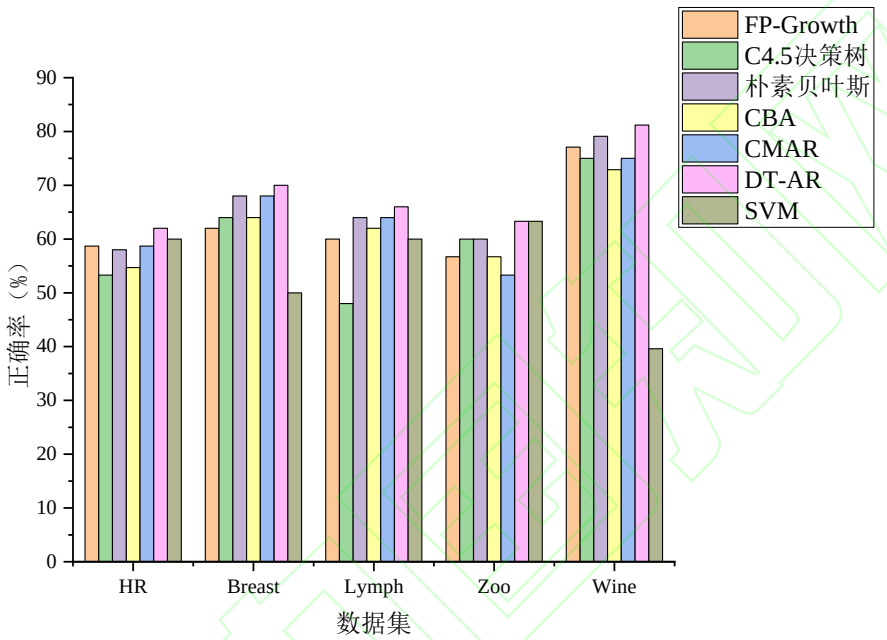

图 6　预测正确率对比图

Fig.6　The comparison of prediction accuracy

将 FP-Growth 算法、CBA 算法、CMAR 算法以及 DT-AR 算法得到的关联规则集在训练集和预测集上的分类精度情况总结如表 8 所示。对比可知，所有关联规则集在训练集上的分类精度均明显高于预测集，说明上述算法均存在过拟合的问题，但 DT-AR 算法的关联规则集在训练集和预测集之间的分类精度差距普遍较小，这说明改进算法在一定程度上改善了规则过拟合的问题。

表 8 不同算法训练集和预测集分类精度对比

Table8 Thecomparison of classification accuracy between training sets and prediction setby different algorithms

方法	指标	HR	Breast	Lymph	Zoo	Wine
FP-Growth	训练集精度	66.7%	74.2%	75.0%	67.6%	78.5%
	预测集精度	58.7%	62.0%	60.0%	56.7%	77.1%
	偏差	12.0%	12.2%	15.0%	10.9%	1.4%
DT-AR	训练集精度	68.0%	75.0%	76.1%	73.2%	82.3%
	预测集精度	62.0%	70.0%	66.0%	63.3%	81.2%
	偏差	6.0%	5.0%	10.1%	9.9%	1.1%
CBA	训练集精度	79.2%	76.7%	83.7%	64.8%	86.1%
	预测集精度	54.7%	64.0%	62.0%	56.7%	72.9%
	偏差	24.5%	12.7%	21.7%	8.1%	13.2%
CMAR	训练集精度	75.2%	80.1%	76.1%	71.8%	83.1%
	预测集精度	58.7%	68.0%	64.0%	53.3%	75.0%
	偏差	16.5%	12.1%	12.1%	18.5%	8.1%

从实验对比结果中可以发现，DT-AR 算法和 CMAR 算法虽然在训练集与预测集分类精度偏差上存在明显差距，但在分类准确率的性能指标上较为接近，没有明显的差距，这是由于 CMAR 算法在生成规则时同样对模式进行了剪枝操作。CMAR 算法的具体流程为：首先，采用 FP-Growth 算法的原理挖掘满足最小支持度和置信度阈值的规则集，并利用一种加强的 FP-树储存规则；然后利用 FP-树提取规则，并根据置信度、相关度和数据覆盖率对规则剪枝，提高规则集的泛化能力；最后利用规则集对数据分类，在面临规则冲突时采取卡方度量确定规则匹配结果。

综上，CMAR 算法在生成规则时同样结合了关联规则挖掘和决策树剪枝的原理，并引入卡方度量选择规则，所以相比于 CBA、FP-Growth、C4.5、SVM 以及朴素贝叶斯分类算法，CMAR 算法通常具备更好的分类效果。但是 CMAR 并未考虑规则过拟合的问题，故在训练集与预测集的分类精度偏差仍然较大，规则过拟合问题仍然较为严重，而 DT-AR 算法不仅结合关联规则与决策树剪枝的优点，具备更高的分类准确率，且通过加入迭代剪枝的方法进一步改善了规则过拟合的问题。

4 结束语

本文提出了一种决策树剪枝加强的关联规则提取方法 DT-AR，在关联规则挖掘 FP-Growth 算法的基础上，利用决策树构建的后剪枝策略作为辅助，补充了关联规则集，并采用置信度作为规则权重系数，优化了规则提取过程，最后对该方法进行了验证分析，实验结果表明经此方法得到的关联规则在面临未知结果的数据集时有更准确的预测结果，且能够得到更为完整的规则集，证明了该关联规则提取方法的有效性。

参考文献：

[1] Agrawal R, Imielinski T, Swami A. Database mining: a performance perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 1993(NO.6):914-925.

[2] 周志华.机器学习[M].No.27,北京:清华大学出版社,2016:75-83.
Zhou Z H. Maching Learning[M].No.27,Beijing:Tsinghua University Press,2016:75-83.

[3] Liu B, Hsu W, Ma Y. Integrating classification and association rule mining[C]//Knowledge Discovery and Data Mining,New York,1998:80-86.

[4] Sivanantham S, Mohanraj V, Suresh Y, et al. Rule precision index classifier: an associative classifier with a novel pruning measure for intrusion detection[J].Personal and Ubiquitous Computing, 2021:1-9.

[5] Jm A, Bka B. A compact and understandable associative classifier based on overall coverage[J]. Procedia Computer Science, 2020,170:1161-1167.

[6] Mabu S, Gotoh S, Obayashi M, et al. A random-forests-based classifier using class association rules and its application to an intrusion detection system[J]. Artificial Life & Robotics, 2016,21(3):371-377.

[7] 王玲, 李树林, 吴璐璐. 基于定量关联规则树的分类及回归预测算法[J]. 工程科学学报, 2016:886-892.
Wang L,Li S L,Wu L L. Categorization and regression algorithm based on the quantitative association rule tree[J].Chinese Journal of Engineering,2016:886-892.

[8] 崔巍, 贾晓琳, 樊帅帅, 等. 一种新的不均衡关联分类算法[J]. 计算机科学, 2020,47(S1):488-493.
Cui W, Jia X L, Fan S S, et al. NewAssociativeClassi-fication Algorithm for Imbalanced Data[J].Computer Science,2020,47(S1):488-493.

[9] 翟悦, 李楠, 于文武. 基于扩展概念格的带约束关联分类规则挖掘方法[J]. 大连交通大学学报, 2021,42(04):88-93.
Zhai Y, Li N, Yu W W. Constrained Association Rule Classification Approach based on Extended Concept Lattice[J]. Journal of Dalian Jiaotong University, 2021, 42(04): 88-93.

[10] 佟玉军, 李煜, 陈文实, 等. 一个改进的分类关联规则挖掘算法[J]. 辽宁工业大学学报（自然科学版）, 2011, 31(5): 287-290.
Tong Y J, Li Y, Chen W S, et al. Improved Class Association Rule Mining Algorithm[J]. Journal of Liaoning University of Technology (Natural Science Edition), 2011, 31(5): 287-290.

[11] Mahdi M A, Hosny K M, Elhenawy I. FR-Tree: A novel rare association rule for big data problem[J].Expert Systems with Applications,2022,187:1-12.

[12] Sanz J, Sesma-Sara M, Bustince H. A fuzzy association rule-based classifier for imbalanced classification problems[J]. Information Sciences, 2021,577(4):266-279.

[13] P. D, M. J, L. P, et al. A novel application of Hoeffding's inequality to decision trees construction for data streams: 2014 International Joint Conference on Neural Networks (IJCNN)[C],Beijing,2014:3324-3330.

[14] 张珏, 田建学, 董婷. 一种基于代价敏感集成决策树的不平衡数据分类方法研究[J]. 榆林学院学报, 2021, 31(02): 53-55.
Zhang Y,Tian J X,Dong T.Research on Cost Senstive Decision Tree for Imbalanced Data[J].Journal of Yulin University,2021,31(02):53-55.

[15] 赵宁杰, 李雪飞. 基于 bagging 思想的决策树分类算法研究[J]. 北京服装学院学报(自然科学版), 2020, 40(03):43-48.
Zhao N J,Li X F.Research on Decision Tree Classification Algorithm Based on the Idea of Bagging[J]. Journal of Beijing Institute of Fashion Technology (Natural Science Edition), 2020,40(03):43-48.

[16] 段化娟, 尉永清, 刘培玉, 等. 一种面向不平衡分类的改进多决策树算法[J]. 广西师范大学学报(自然科学版), 2020,38(02):72-80.
Duan H J, Wei Y Q, Liu P Y, et al. AnImprovedMulti-decisionTreeAlgorithmforImbalancedClassification[J].JournalofGuangxiNormalUniversity (Natural Science Edition), 2020,38(02):72-80.