提取有效规则的关联分类算法

武建华[1,2]，沈钧毅[1]，方加沛[2]
(1. 西安交通大学电子与信息工程学院，710049，西安；2. 暨南大学珠海学院计算机科学系，519070，广东珠海)

摘要：针对关联分类算法产生的规则普遍存在分类器分类精度、效率低的问题，提出了一种提取有效规则的关联分类算法——ACDER算法.首先定义了剩余支持度和剩余置信度，然后通过计算规则剩余支持度和剩余置信度建立了分类器并进行剪枝，以达成对分类尽量少且最有效的规则构成分类器，确保分类器中不存在任何冗余规则和冲突规则.在8个数据集上的测试结果表明，所提算法的平均分类精度比关联规则算法提高了4.15%，而在所有数据源分类器上的规则数却减少了54%.

关键词：关联规则；关联分类；有效规则；分类器
中图分类号：TP301 **文献标志码**：A **文章编号**：0253-987X(2009)04-0022-04

An Associative Classification Algorithm by Distilling Effective Rules

WU Jianhua[1,2]，SHEN Junyi[1]，FANG Jiapei[2]
(1. School of Electronics and Information Engineering，Xi′an Jiaotong University，Xi′an 710049，China；2. Department of Computer Science，Zhuhai Collage，Jinan University，Zhuhai，Guangdong 519070，China)

Abstract：Associative classification algorithms commonly have low efficiency and accuracy. A new associative classification algorithm by distilling effective rules，called ACDER，is presented. Both the remaining support and remaining confidence are defined. Then association classifier is constructed and pruned by distilling the most effective rules to ensure that there exists no any redundant and conflictive rules in the classifier. Experiment results on eight data sets show that the average accuracy of the classifier is 4.15% higher while the average number of rules in classifier is 54% lower than the CBA classification method.

Keywords：association rule；associative classification；effective rule；classifier

数据挖掘中常用的分类方法有决策树、贝叶斯分类和神经网络等，已被广泛、有效地应用于科学实验、医疗诊断、商业预测等领域[1-2].理论上，贝叶斯分类应具有最小的出错率，但其假定的不准确性，使之应用不尽人意.

最近几年提出了一种采用关联规则构造分类器的方法，称之为关联分类方法，它集分类器构造与属性相关分析于一体，是一种具有潜力的分类方法.据此，有关研究相继展开并提出了下述方法：关联聚类系统（ARCS）[3]、多维关联规则的分类算法（CMAR）[4]、关联规则发现方法（CBA）[5]、关联决策树（ADT）[6]等.

由于关联分类方法仅考虑了项目间所有可能的相互关系，因此挖掘出的关联规则可能存在大量的冗余和冲突规则.针对这一问题，本文提出了一种提取有效规则的关联分类算法——ACDER算法，可消除关联分类中所有的冗余规则和冲突规则，提取出的有效规则集可生成一个高精度分类器.

1 基于有效规则提取的关联分类算法

1.1 剩余支持度和剩余置信度

(1)基本概念.为了客观地评价规则在分类过程

收稿日期：2008-07-14. 作者简介：武建华(1963—)，女，博士生；沈钧毅(联系人)，男，教授，博士生导师. 基金项目：国家自然科学基金资助项目(60673087).

中的作用,建立了分类规则度量的 2 个新概念:剩余支持度和剩余置信度,定义如下.

定义 1　有 2 条规则 $A\to C$、$B\to D$,且 B 包含 A. 若假设 $B\to D$ 的置信度较大,并选择 $B\to D$ 规则进行分类,则 $A\to C$ 规则的剩余支持度是指在删除了 $B\to D$ 所覆盖的训练样本对象后,用训练集对 $A\to C$ 规则重新计算支持度得到的值.

定义 2　有 2 条规则 $A\to C$,$B\to D$,且 B 包含 A. 若假设 $B\to D$ 的置信度较大,并选择 $B\to D$ 规则进行分类,则 $A\to C$ 规则的剩余置信度是指在删除了 $B\to D$ 所覆盖的训练样本对象后,用训练集对 $A\to C$ 规则重新计算置信度得到的值.

结论　若一条规则的剩余支持度为 0,则相应的剩余置信度也为 0.

这里,规则所覆盖的训练样本对象是指样本集中满足规则前件的所有训练样本. B 包含 A,说明这 2 条规则不是冗余规则就是冲突规则.

(2)分类规则置信度的变化趋势. 在关联分类中,为了有效阻止与分类规则有冗余和冲突的规则进入分类器,可采用如下思路:①使冗余和冲突规则在分类器构造过程中尽可能向类关联规则集的后部移动;②通过设置一个度量阈值,滤掉尽可能多的冗余和冲突规则. 对于前者可通过分析类关联规则置信度的变化趋势找到解决方法,对于后者可用最小置信度作为度量阈值.

当一条规则被选择用于分类后,为了考虑它对类关联规则集中剩余规则的影响,可对这些剩余规则重新计算置信度,即计算剩余置信度,以真实地反映这些规则在分类中的作用. 如果这样,剩余规则的置信度在每选一条规则加入分类器后都有可能发生变化,其变化情况有以下几种:①若用于分类的规则与某条剩余规则不冗余或不冲突,并且两规则覆盖的样本对象不相交,则该条剩余规则的置信度不变化;②若用于分类的规则与某条剩余规则不冗余或不冲突,但两规则覆盖的样本对象部分相交,则该条剩余规则的置信度变化的大小不确定;③若用于分类的规则与某条剩余规则是冗余或冲突的,并且两规则覆盖的样本对象完全相同,则该条剩余规则的剩余置信度等于 0;④若用于分类的规则与某条剩余规则是冗余的,并且两规则覆盖的样本对象不完全重叠,则该条剩余规则的剩余置信度为 0 或减小;⑤若用于分类的规则与某条剩余规则是冲突的,并且两规则覆盖的样本对象不完全重叠,则该条剩余规则的剩余置信度为 0 或变化的大小不确定.

情况① ~情况③ 是显然的,所以仅需证明后 2 种情况.

证明④　已知 $A\to C$ 和 $B\to D$ 是两条冗余规则,并且两规则覆盖的样本对象不完全相同,设 B 包含 A,则 $A\to C$ 覆盖的样本对象包含 $B\to D$ 覆盖的样本对象. 若使用 $A\to C$ 进行分类,删除 $A\to C$ 规则所覆盖的样本对象后,$B\to D$ 规则覆盖的训练样本对象为空,所以其剩余支持度为 0,由定义 2 知剩余置信度为 0;若使用 $B\to D$ 进行分类,当两条规则是冗余规则时,C 和 D 是相同的类标识,均记为 C,则 $A\to C$ 的置信度变化

$$\Delta d_{\mathrm{conf}}=d_{\mathrm{conf}}(A\to C)-d_{\mathrm{re\_conf}}(A\to C)\quad(1)$$

式中:$d_{\mathrm{re\_conf}}(A\to C)$ 表示删除覆盖样本后的剩余置信度. 在这种情况下,由剩余置信度的定义可知

$$d_{\mathrm{re\_conf}}(A\to C)=\frac{d_{\mathrm{sup}}(A\to C)-d_{\mathrm{sup}}(B\to C)}{d_{\mathrm{sup}}(A)-d_{\mathrm{sup}}(B)}\quad(2)$$

代入式(1)得

$$\Delta d_{\mathrm{conf}}=d_{\mathrm{conf}}(A\to C)-\frac{d_{\mathrm{sup}}(A\to C)-d_{\mathrm{sup}}(B\to C)}{d_{\mathrm{sup}}(A)-d_{\mathrm{sup}}(B)}=$$

$$d_{\mathrm{conf}}(A\to C)-\frac{d_{\mathrm{conf}}(A\to C)-d_{\mathrm{sup}}(B\to C)/d_{\mathrm{sup}}(A)}{1-d_{\mathrm{sup}}(B)/d_{\mathrm{sup}}(A)}$$

记 $\alpha=d_{\mathrm{sup}}(B)/d_{\mathrm{sup}}(A)$,并代入式(2),最后可得

$$\frac{d_{\mathrm{sup}}(B\to C)-d_{\mathrm{conf}}(A\to C)d_{\mathrm{sup}}(B)}{d_{\mathrm{sup}}(A)-d_{\mathrm{sup}}(B)}=\frac{d_{\mathrm{conf}}(B\to C)-d_{\mathrm{conf}}(A\to C)}{(d_{\mathrm{sup}}(A)/d_{\mathrm{sup}}(B))-1}\quad(3)$$

因为 B 包含 A,所以 $B\to C$ 覆盖的样本对象是 $A\to C$ 覆盖的样本对象的子集,所以 $d_{\mathrm{sup}}(A)>d_{\mathrm{sup}}(B)$. 又假设,首先选择置信度高的规则用于分类,则 $d_{\mathrm{conf}}(B\to C)>d_{\mathrm{conf}}(A\to C)$,所以 $\Delta d_{\mathrm{conf}}>0$,即规则的剩余置信度朝着递减的方向变化.

证明⑤　已知 $A\to C$ 和 $B\to D$ 是 2 条冲突规则,该规则覆盖的样本对象不完全相同. 设 B 包含 A,则 $A\to C$ 覆盖的样本对象包含 $B\to D$ 覆盖的样本对象. 若使用 $A\to C$ 进行分类,在删除了 $A\to C$ 规则所覆盖的样本对象后,$B\to D$ 规则覆盖的训练样本对象为空,所以其剩余支持度为 0,由定义 2 知剩余置信度为 0;若使用 $B\to D$ 进行分类,在这种情况下式(2)仍然成立,其证明过程与情况④类似. 但是,$d_{\mathrm{conf}}(B\to D)>d_{\mathrm{conf}}(A\to C)$,不能确定 $d_{\mathrm{conf}}(B\to C)>d_{\mathrm{conf}}(A\to C)$,故 Δd_{conf} 的变化无法确定.

综合上述讨论,在③、④以及部分⑤的情况下,随着可选规则不断加入分类器,剩余置信度减小. 因

此，删除剩余置信度小于最小置信度阈值的规则，将大大减少冗余规则和冲突规则进入分类器.

1.2 基于有效规则提取的关联分类算法

该算法由 3 部分组成：①根据关联规则算法发现所有类关联规则，这与 CBA 算法的第一步相同；②通过重新计算规则剩余支持度和剩余置信度建立分类器并剪枝；③利用分类器对未知类别数据进行分类.

算法步骤为：①应用带约束的关联规则产生所有满足最小支持度和最小置信度的类关联规则；②将类关联规则集按置信度、支持度从大到小和短规则优先等排序；③选择排在最前面的规则加入分类器，同时将该规则和所覆盖的所有训练样本对象从训练样本集中删除；④基于剩余的训练样本，重新计算剩余规则的剩余支持度和剩余置信度；⑤按校正后的置信度重新从大到小排序，删除置信度小于最小置信度阈值的规则；⑥重复步骤③～步骤⑤直到没有剩下任何规则可用，停止算法的循环.

算法的形式化描述如图 1 所示. 另外，计算所有不满足分类器中规则的样本在各类上所占的比率，并选择比率最高的类作为默认类.

用本文算法构造的分类器已消除了大量冗余和冲突规则. 剪枝过程主要分 3 步进行：①假设分类器中同时存在冗余规则 $A\rightarrow C$ 和 $B\rightarrow C$，且 B 包含 A，将规则 $B\rightarrow C$ 从分类器中剪掉，因为满足 B 条件的样本均满足 A 条件，所以去掉 $B\rightarrow C$ 仍能正确分类；②假设分类器中同时存在冲突规则 $A\rightarrow C$ 和 $B\rightarrow D$，且 B 包含 A，将置信度低的规则去掉；③将所有与默认规则具有同样分类标签的规则从分类器中剪掉，因为分类器会应用默认规则将那些本来满足被剪规则的样本分到相同的类中，所以不会影响分类精度.

本文算法的复杂度约为 $O(np^2)$，主要取决于规则的数量 p、训练集样本数量 n，而规则的数量受训练集中样本数量和各种阈值限制的影响.

```
输入：训练集D，最小支持度阈值mind'sup，最小置
      信度阈值mind'conf
输出：分类器c
1: c=∅;
2:Strong_Rules=genRules(D,mind'sup,mind'conf);
3: CARs=Class_Strong_Rules();
4: Rules_Set=Sort_ CARs (); //按置信度、支持度、
   短规则排序
5: for Rules_Set <>∅ do
6:   r=first_rule;
7:   k=Delete_Cover_Simplesnumber();
8:   if (k<>0) then
9:     c =∪ {r};
10:  end if
11:  delete r from Rules_Set;
12:  for each r∈Rules_Set do
13:    dsup=dre_sup();
14:    if (dsup ==0 ) then
15:      dconf=0;
16:    else
17:      dconf = dre_sup ();
18:    end if
19:    if dconf <min dconf then
20:      delete r from Ruels_Set;
21:    end if
22:  end for
23:  Sort_Rules_Set (); // 按置信度由大到小排序
24: end for
25: cut();
26: output: C
```

图 1 基于有效规则提取的关联分类算法

2 实验结果

2.1 实验设置

(1)测试数据. 本文选用了 UCI ML Repository[7]标准数据中的 8 个数据集(见表 1)进行实验，类别数分别为 2、3、4，并对数据集中的连续属性做离散化处理.

(2)验证方法. 8 个数据集中的 5 个给定了一个数据集，剩余的 3 个给定了训练集和测试集. 对给定的数据集采用 5 折交叉法验证，用训练集产生分类器，用测试集检验分类精度.

表 1 实验数据集

数据集	离散性属性数	连续性属性数	样本数	类别数
monks-1	7	0	124	2
monks-2	7	0	169	2
monks-3	7	0	122	2
car	7	0	1 728	4
bupa	1	6	345	2
pima-indians-diabetes	1	8	768	2
new-throid	1	5	215	3
tetris	5	7	1 000	3

(3)度量标准. 估计分类法的准确率就是估计一个给定的分类算法对未来的数据(即未经分类法处理的数据)正确标号的准确率(精度)，即

$$\text{准确率(精度)} = \frac{\text{正确预测的实例数}}{\text{预测的总实例数}}$$

2.2 实验结果

对从 UCI 机器学习仓库获得的 8 个数据集进行实验，每个数据集的最小支持度和最小置信度如表 2 所示.

实验结果如表 3 和表 4 所示，从表中可以看出：本文方法在 5 个数据集中分类精度大于 CBA，在 8 个数据集上的平均分类精度比CBA提高4.15%；

表 2　数据集最小支持度和最小置信度

数据集	最小支持度/%	最小置信度/%
monks-1	2	60
monks-2	2	60
monks-3	2	60
car	1	60
bupa	5	60
pima-indians-diabetes	5	60
new-thyroid	5	60
tetris	5	60

在所有数据源中,分类器的规则数均比 CBA 的规则数少,平均值小于 54%以上. 可见,基于有效规则提取的关联分类算法的分类精度比 CBA 更高,分类器规则数更少,因而分类效率提高.

表 3　分类精度实验结果

数据源	分类精度/%		
	本文方法	CBA	C4.5
monks-1	100.000 0	89.607 4	75.694 4
monks-2	79.861 1	74.074 1	65.046 3
monks-3	92.129 6	95.138 9	97.222 2
car	81.318 6	70.270 7	91.550 9
bupa	64.725 3	66.505 4	69.855 1
pima-indians-diabetes	74.666 7	73.830 8	75.651 0
new-thyroid	81.395 3	81.860 4	94.418 6
tetris	75.500 0	64.000 0	81.000 0
平均值	81.199 6	76.911 0	81.304 8

表 4　分类器大小(规则数)实验结果

数据集	分类器大小		
	本文方法	CBA	C4.5
monks-1	5.000 0	35.000 0	12.000 0
monks-2	53.000 0	79.000 0	20.000 0
monks-3	27.000 0	39.000 0	9.000 0
car	29.400 0	46.400 0	131.000 0
bupa	12.800 0	27.600 0	15.000 0
pima-indians-diabetes	36.600 0	86.800 0	24.000 0
new-thyroid	5.400 0	16.200 0	12.000 0
tetris	22.400 0	91.800 0	21.000 0
平均值	23.950 0	52.725 0	30.500 0

本文方法虽然在 8 个数据集的平均分类精度与著名的 C4.5 分类算法相当,但平均分类器大小比 C4.5 更具优势. monks-3 数据集中含有 16%的噪声数据,此时本文方法的分类精度低于其他 2 种方法,说明它的抗噪能力有所欠缺.

3　结　论

本文提出一种基于有效规则提取的关联分类算法,解决了一般关联分类算法中存在的冗余和冲突规则问题. 与 CBA 相比,不仅分类精度更高,而且分类器中规则数目更少,因而分类效率提高;与 C4.5 相比,虽然在平均分类精度上相当,但平均分类器大小更具优势. 如果考虑其他规则度量标准或权值,可进一步提高分类的准确率,这也是下一步的研究重点.

参考文献:

[1] CHEN Yongqiang, HU Leifang. Study on data mining application in CRM system based on insurance trade [C]// The 7th International Conference on Electronic Commerce. New York, USA: ACM, 2005: 839-841.

[2] 裴晓梅,郑崇勋,徐进. 基于信息积累技术的大脑运动意识任务分类[J]. 西安交通大学学报, 2008,42(6): 756-759.
PEI Xiaomei, ZHENG Chongxun, XU Jin. Classification of motor imagery tasks with information accumulated technique[J]. Journal of Xi′an Jiaotong University, 2008,42(6):756-759.

[3] HANA J W, NISHIOB S, KAWANOC H, et al. Generalization-based data mining in object-oriented databases using an object-cube model[J]. Data and Knowledge Engineering, 1998, 25(1/2): 55-97.

[4] LI Wenmin, HAN Jiawei, PEI Jian. CMAR: accurate and efficient classification based on multiple class-association rules[C]// The 2001 IEEE International Conference on Data Mining. Piscataway, NJ, USA: IEEE, 2001: 369-376.

[5] LIU Bing, HSU W, MA Yiming. Integrating classification and association rule mining [C]// The 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 1998: 80-86.

[6] WANG Ke, ZHOU Senqiang, HE Yu. Growing decision tree on support-less association rules [C]// The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2000: 265-269.

[7] MURPHY-MERZ J C. UCI repository of machine learning databases[EB/OL]. [2008-04-12]. http://www.ics.uci.edu/mlearn/_MLRepository.html.

(编辑　苗凌)