基于关联规则的特征选择算法*

武建华[1,2] 宋擒豹[1] 沈均毅[1] 谢建文[2]

[1](西安交通大学 电子与信息工程学院 西安 710049)
[2](暨南大学 珠海学院 计算机科学系 珠海 519070)

摘 要 关联规则能够发现数据库中属性之间的关联,通过优先选择短规则用于相关属性的选择,有可能得到最小的属性子集. 基于此,本文提出一种基于关联规则的特征选择算法,实验结果表明在属性子集大小和分类精度上优于多种特征选择方法. 同时,对支持度和置信度对算法效果的影响进行探索,结果表明高的支持度和置信度并不导致高的分类精度和小的特征子集,而充足的规则数是基于关联规则特征选择算法高效的必要条件.

关键词 特征选择,特征子集,关联规则,分类
中图法分类号 TP 391

Feature Selection Algorithm Based on Association Rules

WU Jian-hua[1,2], SONG Qin-Bao[1], SHEN Jun-Yi[1], XIE Jian-Wen[2]
[1](*School of Electronic and Information Engineering*, *Xi'an Jiaotong University*, *Xi'an* 710049)
[2](*Department of Computer Science*, *Zhuhai Collage*, *Jinan University*, *Zhuhai* 519070)

ABSTRACT

A feature selection algorithm based on association rules is presented, and the impact of support and confidence on the presented method are studied. The experimental results show that the feature subset size and classification accuracy of the presented method are better than those of other methods. Furthermore, the results indicate high support and confidence levels do not guarantee high classification accuracy and small feature subset, and the sufficient number of rules is the precondition for high efficiency of feature selection based on association rules.

Key Words Feature Selection, Feature Subset, Association Rules, Classification

* 国家自然科学基金资助项目(No. 60673124,60673087)
收稿日期:2008-06-25;修回日期:2008-11-19
作者简介 **武建华**,女,1963 年生,副教授,主要研究方向为数据挖掘. E-mail: tjhwu@ jnu. edu. cn. **宋擒豹**,男,1966 年生,教授,博士生导师,主要研究方向为数据挖掘、软件工程. **沈均毅**,男,1939 年生,教授,博士生导师,主要研究方向为数据挖掘、数据库理论. **谢建文**,男,1985 年生,学士.

1　引　言

特征选择是通过删除不相关属性,能够为特定的应用在不失去数据原有价值的基础上选择最小的属性子集. John[1]认为特征选择是在不降低分类精度条件下,降低特征维数的过程. Koller[2]定义特征选择为在保证结果类分布尽可能与原始数据类分布相似条件下,选择尽可能小的特征子集. 特征选择在模式识别中扮演重要角色,同时也是知识发现预处理过程中重要的一环. 通过特征选择不仅提高了数据质量,加快挖掘过程的速度,而且使得挖掘出的知识更易理解.

Dash[3]等人对数据挖掘领域中特征选择问题进行全面综述. 特征选择算法在特征子集空间中进行搜索时一般要考虑搜索方向、搜索策略、评价方法和停止标准4个方面. 搜索方向即为特征子集产生的次序,有前向搜索、后向搜索、双向搜索和随机搜索. 搜索策略根据搜索空间大小的不同分为完全搜索、启发式搜索和定性搜索. 评价方法决定进一步的搜索方向,通常的评价方法有距离、信息增益、依赖性、一致性和准确性等. 停止标准与搜索算法、评价方法及具体的应用相联,常用的停止标准有循环次数或执行时间、特征数目阈值、评价函数阈值等. 另外,根据样本中是否含有类别信息,特征选择可分为有监督特征选择和无监督特征选择. 有监督特征选择是指在给定类别的前提下,利用特征之间和特征与类别之间的关系对特征集进行选择的过程. 无监督特征选择是指在数据集中,通过数据集中特征自身之间的关系进行特征选择. 特征选择按照和后续分类算法的结合方式可分为嵌入式(Embedded)、过滤式(Filter)、封装式(Wrapper)和组合式(combination)4种. 在嵌入式结构中,特征选择算法本身作为组成部分嵌入到分类算法里. 最典型的是决策树算法. 算法在每一结点选择分类能力最强的特征,然后基于选中的特征进行子空间分割,继续此过程,直到满足终止条件. 可见决策树生成的过程也就是特征选择的过程. 过滤式特征选择的评估标准直接由数据集求得,独立于分类算法,具有通用性强、算法复杂性低、运行效率高、适用于大规模数据集等优点. 但由于忽略了所选特征子集在分类算法上的性能,因此选出的特征子集也许不是最优. 经典的过滤式特征选择算法有 Relief 系列算法等. 封装式特征选择算法最早由 John 等人在 1994 年提出[1],其将分类算法的性能作为特征选择的评估标准. 封装式特征选择算法比过滤式特征选择算法准确率高,但算法效率较低. 通过对两种或以上的常规特征选择算法进行组合、以达到优势互补来进行特征选择的方法,称为组合式特征选择算法.

经典的决策树算法有 Quinlan 的 ID3 和 C4.5 以及 Breiman 的分类和回归树(Classification and Regression Trees, CART)算法等. ID3 使用信息增益作为属性选择度量,该度量基于 Shannon 在研究消息的值或"信息内容"的信息论方面的先驱工作,缺点是倾向于选择具有大量值的属性. C4.5 使用称作增益率的信息增益扩充,克服 ID3 的偏依性,使用"分裂信息"值将信息增益规范化,但倾向于不平衡的分裂,其中一个划分会比其他划分小得多. CART 使用 Gini 指标作为属性选择度量,该指标偏向于多值属性,并且当类的数量很大时会有困难,此外还倾向于导致相等大小的划分和纯度. 尽管上述的决策树方法有缺陷,但在实践中产生相当好的效果.

Kira 和 Rendell 于 1992 年提出一种有效的特性选择算法即 Relief 算法[4],该算法的特征选择标准是特性相关性. Relief 算法最大的局限性是在相关特征集中不能识别出冗余特征,而且一般只能使用二值类别数据. Kononenlo[5]扩展了原始 Relief 算法,使它能够处理多类别、不完整和有燥声的数据.

遗传算法是美国 Michigan 大学 Holland 根据生物进化论和遗传学的思想提出的一种全局启发式优化算法[6]. 它利用遗传算子(选择、交叉和变异),促进解集合类似生物种群在自然界中自然选择、优胜劣汰、不断进化,最终收敛于最优状态. 最早采用遗传算法进行特征选择是 Siedlecki 和 Sklansky[7]. 文献[8]中 Vafaie 等人采用遗传算法进行特征选择,取得较好结果. 遗传算法对问题依赖性小,搜索能力强,适合大规模复杂问题的优化. 但据文献[9]中 Jain 的实验,遗传算法容易过早收敛,需就此采取措施,且当应用于大规模数据,如果采用遗传算法的封装式特征选择,运行效率较低[10].

近年来虽然基于智能计算的特征选择方法受到关注[11-14],比如文献[14]中 Wiratunga 等学者探索了在文本数据中使用聚类和贪心算法进行无监督特征选择的方法,但基于关联规则的特征选择方法研究并不多见. Wiratunga 的另一篇论文[15]给出了在文本实例检索中,通过使用增强的决策树算法进行特征选择,使用关联规则进行特征概化的方法,显著提高检索的准确率.

不同于上述算法,本文利用关联规则算法进行属性选择. 其基本思想是首先挖掘后件为类属性的强关联规则,再根据规则长度、置信度、支持度和提

升度找出与类属性密切相关的属性子集. 实验结果表明在属性子集大小和分类精度上本文方法具有较大优势.

2 基于关联规则的特征选择算法

2.1 关联规则挖掘简介

关联规则挖掘算法首先由 Agrawal 等人[16]在 1993 年提出,其目的是要自动发现数据项之间存在的隐含关联关系.

设 $I=\{i_1,i_2,\cdots,i_m\}$ 是项的集合. 设任务相关的数据 D 是数据库事务的集合,其中每个事务 T 是项的集合,使得 $T\subseteq I$. 每个事务有一个标识符,称作 *TID*. 设 A 是一个项集,事务 T 包含 A 当且仅当 $A\subseteq T$. 关联规则是形如 $A\Rightarrow B$ 的蕴涵式,其中 $A\subset I$,$B\subset I$,并且 $A\cap B=\varnothing$.

关联规则 $A\Rightarrow B$ 可用如下参数描述.

1) 支持度.

$$Support(A\Rightarrow B)=P(A\cup B).$$

2) 置信度.

$$Confidence(A\Rightarrow B)=P(B\mid A)=\frac{sup(A\cup B)}{sup(A)}.$$

3) 提升度.

$$Lift(A\Rightarrow B)=\frac{sup(A\cup B)}{sup(A)\times sup(B)}.$$

Support 是 D 中事务同时包含 A 和 B 二者的百分比,是对关联规则重要性的衡量,说明该规则在所有的事物中有多大的代表性. 如果项集满足最小支持度 *min_sup*,则称它为频繁项集. *Confidence* 是指 D 中包含 A 的事务同时也包含 B 的百分比,是对关联规则准确度的衡量. 它反映在给定 A 的前提下,B 发生的后验概率. *Lift* 有时也称为 *Interest*,它是 A 和 B 同时发生的概率和在假定 A 和 B 独立的前提下 A 和 B 同时发生概率之间的比值. *Lift* 用来衡量 A 和 B 之间的关联与 A 和 B 相互独立偏离的程度. 如果 *Lift* 接近于 1,A 和 B 就是独立的;如果 *Lift* 小于 1,A 和 B 互为抑制;*Lift* 越大于 1,规则的实际意义就越好.

2.2 基于关联规则的特征选择算法

虽然特征选择方法已在模式识别领域和数据挖掘得到充分研究,但基于关联规则的特征选择方法研究并不多见. 由于关联规则能够发现数据库中属性之间的关联,通过优先选择短规则用于相关属性的选择,有可能得到最小的属性子集. 因此,本文提出一种基于关联规则的特征选择方法. 实验结果表明在属性子集大小和分类精度上优于多种特征选择方法.

该方法主要包括 3 个阶段:生成规则集,构造属性集,测试属性集.

第一阶段应用 Apriori 算法从训练集中生成所有后件为类属性且提升度大于 1 的规则. 生成关联规则时,支持度和置信度由用户指定.

第二阶段是通过给定的循环次数从规则集中选出规则,并将规则前件的属性添加到属性集中,循环结束后属性集中的所有属性就是特征选择的结果. 而循环次数的确定是本文方法要解决的一个重点. 循环次数的确定必须满足两个条件:1) 选择出的属性子集小;2) 按属性子集中的属性进行分类的精度高. 在本文方法中,通过在多个数据集上实验探索,找出循环次数和属性子集大小及分类精度之间的联系,从而较好地解决了这一问题. 详见实验部分.

第三阶段是测试阶段,用于评估特征选择的效果,本文使用 C4.5 分类器对特征选择处理后的数据集进行分类,分类正确率作为衡量特征子集好坏的一个指标.

第一、二阶段对应算法的主要步骤如下.

step 1 使用 Apriori 算法对训练集产生强关联规则集,如果训练集的非类属性为连续型属性,对其进行离散化.

step 2 从规则集中挑选所有后件为类属性的规则,形成新规则集并计算规则集中每条规则的提升度,删除提升度小于 1.0 的规则,形成有效规则集.

step 3 对有效规则集按照长度最短、置信度最大、支持度最大进行多关键字排序.

step 4 若循环次数大于给定的阈值则输出属性集中的属性并退出. 否则,从有效规则集中取出第一条规则,将其前件的属性添加到属性集合里,并从有效规则集中删除该条规则. 注意若前件中的某个属性已在属性集合中存在,则不再添加.

step 5 从训练集中将该条规则所覆盖的样本删除,并对剩余训练集重新计算有效规则集中剩余规则的置信度和支持度.

step 6 对有效规则集按照置信度最大、支持度最大排序,继续循环转 step 4. 其形式化描述如下.

算法 基于关联规则的特征选择算法

输入 训练集 D,最小支持度阈值 *min_sup*,最小置信度阈值 *min_conf*,类属性名 *class*,循环次数阈值 k

输出 特征集 *result*

```
1  result = ∅;
2  R = genRules(D,min_sup,minconf,class);
```

```
3   sort1(R);
4   for (int i = 1; i ≤ k; i ++) do
5      if R = = Ø then break;
6      else r = remove-first(R);
7         S = get-features-in-premise(r);
8         result = ∪ S;
9         D_delete = filer-trainset(r, S);
10          reset(D_delete, R);
11          sort2(R);
12       end if
13  end for
14  output: result;
```

下面分析算法的时间复杂度. 对于第 2 步产生所有有意义的带约束关联规则,目前已经有很多关联规则挖掘算法,如经典的 Apriori 及其改进算法,以及 FP-增长算法等,在这些算法上稍加修改,很容易使产生的规则都有一个类标号,因此不再关注这一步的时间复杂度. 而重点来分析根据已产生规则进行特征选择的时间复杂度. 设训练集有 m 个特征,n 个样本,第 2 步共挖掘出 p 条规则,则第 3 步和第 11 步多关键字排序的时间为 p^2,第 5 步到第 8 步的计算时间各为 1,第 9 步计算时间为 mn,第 10 步计算时间为 np,循环 k 次(k 取值为 3 或 6,见实验部分). 这样,k 次循环所需的总时间为 $k(5 + mn + np + p^2)$. 因此,根据规则的特征选择算法的时间复杂度为 $O(mn + np + p^2)$. 由此可见,该算法的时间复杂度主要取决于规则的数量 p、训练集特征数量 m 和样本数量 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.

3　实验结果与分析

3.1　实验设置

1) 实验环境. 本文方法用 Java 编程实现,分类实验和其它特征选择方法实验均使用怀卡托智能分析环境(Waikato Environment for Knowledge Analysis) 即 Weka 软件平台完成.

2) 测试数据. 为了测试基于关联规则的特征选择算法的性能,本文选用数据挖掘领域用来对比不同算法性能的标准数据 UCI ML Repository[17] 中的 10 个数据集进行实验. 数据类型不同(只包含离散型或连续型特征的单一类型以及同时包含离散型和连续型特征的混合类型) 和有的数据集存在缺失数据使得数据有较广泛的代表性,可以较好地验证特征选择方法在大规模实际数据集上的性能. 表 1 给出选用 10 个数据集的详细信息.

3) 验证方法. 10 个数据集中有 6 个数据集只给定 1 个数据集合,4 个数据集分别给定训练集和测试集. 为了比较本文方法在一个数据集上使用前后的效果,对 10 个数据集均采用 10-折交叉验证,将数据集分成 10 份 $S_1, S_2, \cdots, S_{10}$,训练和测试进行 10 次,在第 i 次迭代,S_i 用作测试集,其余子集都用于训练分类算法. 准确率估计是 10 次迭代正确分类数除以初始数据中的样本总数. 其中 4 个数据集只用了训练集.

4) 度量标准. 正确分类的实例是衡量一个分类算法性能最重要的标准. 估计分类法的准确率使得我们可以估计一个给定的分类算法对未来数据(即未经分类法处理的数据) 正确标号的准确率,其定义如下:

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

本文选用准确率和特征子集大小作为算法性能的度量.

5) 基准方法. 为了更好地分析本文方法的准确率,本文将其与使用本文方法前及常用的 3 种特征选择方法进行对比. 通过分别在 10 个数据集上运行这 4 类算法,测试它们分类的准确率和特征子集大小,从而说明本文方法的有效性.

表 1　实验数据集

Table 1　Experimental datasets

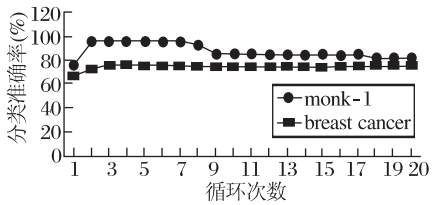
数据集	离散型特征数	连续型特征数	缺失数据	样本数	类别数
ionosphere	0	34	n	351	2
vote	16	0	y	435	2
wine	0	13	n	178	3
horse-clonic	20	7	y	300	2
zoo	17	0	n	101	7
breast-cancer	0	10	y	699	2
monk-1	7	0	n	124	2
monk-2	7	0	n	169	2
monk-3	7	0	n	122	2
mushroom	22	0	y	8124	2

3.2　循环次数的确定

循环次数是基于关联规则特征选择算法的一个重要参数,直接影响特征选择的个数和效果. 实验数据集有 ionosphere、wine、zoo、monk-1、breast-cancer. 通过在 5 个数据集中以不同的循环次数参数执行基于关联规则的特征选择算法,记录每次实验后所选的特征,并使用这些特征对数据集样本进行分类,得到分类准确率. 图 1 是 5 个数据集在使用基于关联规则的特征选择算法进行特征选择时,循环次数取值 1 ~ 20 的分类准确率. 由图 1(a) 可以发现,对特征

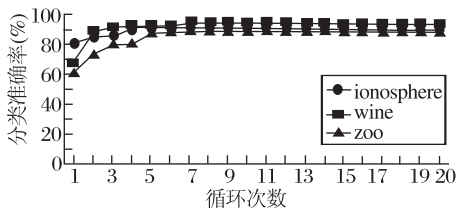
个数小于等于10的monk-1和breast cancer数据集，在循环次数为3时，分类准确率达到最大值. 由(b)可以看出，对特征个数大于10的3个数据集，在循环次数为6时，有2个数据集分类准确率达到最大值，1个数据集的分类准确率最接近最大值.

图1(c)是5个数据集在使用基于关联规则的特征选择算法进行特征选择时，循环次数取值1～20时选择的特征占总特征的比例. 对特征个数小于等于10的monk-1和breast cancer数据集，在循环次数为3时，特征比例维持在一个较小值. 对特征个数大于10的其余3个数据集，在循环次数为6时，特征比例也维持在一个较小值.



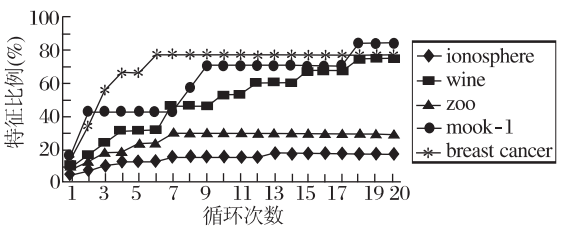
(a) 属性小于等于10的数据集不同循环次数下的分类准确率

(a) Classification accuracy with different cyclic number and attributes ≤ 10



(b) 属性大于10的数据集不同循环次数下的分类准确率

(b) Classification accuracy with different cyclic number and attributes > 10



(c) 不同循环次数下的特征比例

(c) Feature proportion with different cyclic number

图1 循环次数的确定

Fig. 1 Selection of cyclic number

一个好的特征选择算法应当是特征数目少但分类精度高. 因此，综合实验数据得到以下结论：基于关联规则的特征选择算法，对于特征数小于等于10的数据集，循环次数取值为3以保证低特征数，对于特征数大于10的数据集，循环次数设为6以获得良好的综合效果.

3.3 基于关联规则的特征选择算法实验

在weka平台上对每个数据集，运行信息增益特征选择方法、遗传特征选择方法和组合特征选择方法，记录每种方法选择的特征子集. 运行本文方法并记录选择的特征子集. 用weka平台分别对10训练数据集根据特征子集选择相应的特征，使用10-折交叉验证法得到C4.5的分类准确率. 该分类准确率和特征子集的大小是评估特征选择算法性能的两个重要指标.

表2给出在10个数据集中使用本文方法前后在特征子集大小和分类正确率上的比较. 实验结果表明基于关联规则的特征选择算法产生的特征子集与原始特征集相比，特征数目大大减少. 而分类测试准确率总体上比原特征集的分类正确率略有提高，说明基于关联规则的特征选择算法不仅降低了特征维数，而且保证了分类器的精度.

表2 本文方法实验结果

Table 2 Experimental results of proposed method

数据集	全特征		基于关联规则	
	特征数	分类正确率	特征数	分类正确率(支持度,置信度)
ionosphere	34	91.4530	4	91.7379 (30,50)
vote	16	96.3218	3	95.6322 (50,50)
wine	13	93.8202	3	95.5056 (15,50)
horse-clonic	27	81.2709	3	81.2709 (50,50)
zoo	17	92.0792	4	91.0891 (40,50)
breast cancer	10	75.5245	5	75.8741 (20,50)
monk-1	7	82.2581	3	95.9677 (5,50)
monk-2	7	56.2130	5	57.3964 (5,50)
monk-3	7	93.4426	3	93.4426 (5,50)
mushroom	22	100	6	99.1137 (40,50)
平均	16	86.2383	3.9	87.7030

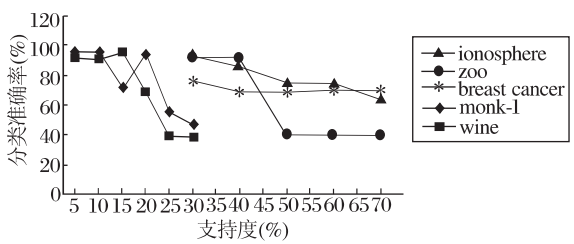
表3对基于关联规则的特征选择算法和过滤式、封装式和组合式的特征选择算法的效果进行比较. 说明基于关联规则的特征选择算法在分类准确率上与其他3种算法不分上下，不同类型的特征选择算法在不同的数据集上各有优势，而基于关联规则的特征选择算法在特征子集大小上优势明显，特征子集数目比其他3种算法小得多，总体效果优于其他3种算法.

表 3　4 种特征选择算法实验结果比较

Table 3　Experimental result comparison among 4 feature selection algorithms

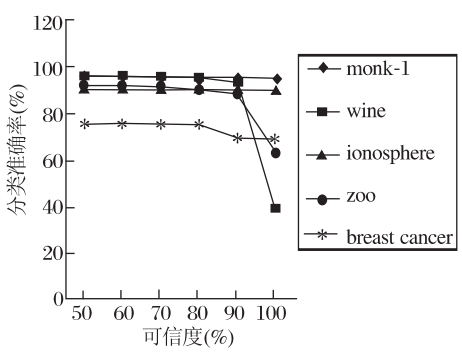
数据集	基于关联规则		过滤式(信息增益)		封装式(遗传算法)		组合式(排序搜索 + Relief)	
	特征数	分类正确率	特征数	分类正确率	特征数	分类正确率	特征数	分类正确率
ionosphere	4	91.7379	33	91.1681	13	85.4701	33	91.4530
vote	3	95.6322	16	96.3218	7	96.3218	13	96.3218
wine	3	95.5056	13	93.8202	5	94.3820	5	94.3820
horse-clonic	3	81.2709	21	80.9365	6	86.6221	2	79.5987
zoo	4	91.0891	17	92.0792	7	93.0693	14	92.0792
breast cancer	5	75.8741	9	75.5245	6	73.0763	3	72.7273
monk-1	3	95.9677	7	82.2581	3	95.9677	3	95.9677
monk-2	5	57.3964	7	56.2130	6	56.2130	6	56.2130
monk-3	3	93.4426	7	93.4426	2	93.4426	2	93.4426
mushroom	6	99.1137	21	100	7	100	8	100
平均	3.9	87.7030	15.1	86.1764	6.2	87.4565	8.9	87.2185

从而得出结论,基于关联规则的特征选择算法在保证分类器精度前提下,大幅度降低数据集的特征数,达到较好的特征约简效果,为后续对数据集进行数据挖掘、模式识别或机器学习提供一个高效的预处理方法.



(a)支持度逐渐变大

(a)Support increasing



(b)置信度逐渐变大

(b)Confidence increasing

图 2　参数对分类准确率的影响

Fig. 2　Impact of parameters on classification accuracy

3.4　参数对算法效果的影响

本 文方法在使用时要设置最小支持度和最小置信度两个参数. 而参数对本文方法效果的影响是需要探索的问题. 图 2 是在 5 个数据集上支持度、置信度取不同值时分类准确率的变化情况. 从图中可以看出,随着支持度逐渐变大,分类准确率有一定起伏,但最终均降低了分类准确率. 随着置信度逐渐变大,分类准确率基本维持在开始值. 但当置信度增加到 90% 左右时,有 3 个数据集分类准确率有明显下降. 表 4、表 5 分别给出在不同支持度和置信度条件下产生的规则条数. 从表中可以发现,随着支持度和置信度的逐渐增大,规则的条数在迅速减少,因而有用的规则数目也在减少. 对比本小节的图与表可以得出结论:不同的支持度和置信度参数设置对算法效果有直接影响,高的支持度和置信度并不导致高的分类精度和小的特征子集,而充足的规则数是基于关联规则的特征选择算法高效的必要条件.

表 4　5 个数据集最小置信度为 50% 的规则条数

Table 4　Number of rules on 5 datasets (minimum confidence = 50%)

最小置信度	monk-1	wine	ionos-phere	zoo	breast cancer
5%	680	421	–	–	–
10%	158	33	–	–	–
15%	54	5	–	–	–
20%	21	2	–	–	–
30%	–	–	818	29623	167
40%	–	–	72	5960	58
50%	–	–	24	1062	51
60%	–	–	12	202	12

表 5 5 个数据集中不同最小支持度的规则条数

Table 5 Number of rules on 5 datasets with different minimum supports

数据集	最小支持度	可信度				
		50%	60%	70%	80%	90%
monk-1	5%	680	326	162	77	47
wine	5%	421	279	190	143	115
ionos phere	30%	818	701	594	393	200
zoo	40%	5960	4985	3824	2572	1563
breast cancer	20%	580	484	397	254	99

4 结束语

本文提出一种基于关联规则的特征选择算法,给出算法的描述过程和实验分析过程. 结果表明本文算法在属性子集大小和分类精度上优于多种特征选择方法. 另外,本文还探索了支持度和置信度对特征选择算法效果的影响,实验结果表明高的支持度和置信度并不导致高的分类精度和小的特征子集,而充足的规则数是基于关联规则的特征选择算法高效的必要条件.

参 考 文 献

[1] John G H, Kohavi R, Pfleger K. Irrelevant Features and the Subset Selection Problem // Proc of the 11th International Conference on Machine Learning. New Brunswick, USA, 1994: 121-129

[2] Koller D, Sahami M. Toward Optimal Feature Selection // Proc of the International Conference on Machine Learning. Bari, Italy, 1996: 284-292

[3] Dash M, Liu H. Feature Selection for Classification. Intelligent Data Analysis, 1997, 1(3): 131-156

[4] Kira K, Rendell L A. The Feature Selection Problem: Traditional Methods and a New Algorithm //Proc of the 9th National Conference on Artificial Intelligence. San Jose, USA, 1991: 129-134

[5] Kononenko I. Estimating Attributes: Analysis and Extension of Relief // Proc of the European Conference on Machine Learning. Catania, Italy, 1994: 171-182

[6] Michalewicz Z. Genetic Algorithms+Data Structures=Evolution Programs. New York, USA: Springer-Verlag, 1996

[7] Siedlecki W, Sklansky J. On Automatic Feature Selection. International Journal of Pattern Recognition and Artificial Intelligence, 1988, 2(2): 197-220

[8] Vafaie H, de Jong K. Genetic Algorithm as a Tool for Feature Selection in Machine Learning //Proc of the 4th International Conference on Tools with Artificial Intelligence. Arlington, USA, 1992: 200-204

[9] Jain A, Zongker D. Feature Selection: Evaluation, Application, and Small Sample Performance. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, 19(2): 153-158

[10] Martin-Bautista M J, Vila M A. A Survey of Genetic Feature Selection in Mining Issues // Proc of the Congress on Evolutionary Computation. Washington, USA, 1999, Ⅱ: 13-23

[11] Waske B, Schiefers S, Braun M. Random Feature Selection for Decision Tree Classification of Multi-Temporal SAR Data // Proc of the IEEE International Geoscience and Remote Sensing Symposium. Denver, USA, 2006: 168-171

[12] Tian D, Keane J, Zeng Xiaojun. Evaluating the Effect of Rough Set Feature Selection on the Performance of Decision Trees // Proc of the IEEE International Conference on Granular Computing. Atlanta, USA, 2006: 57-62

[13] Chen Huanhuan, Yao Xin. Evolutionary Multiobjective Ensemble Learning Based on Bayesian Feature Selection // Proc of the IEEE Congress on Evolutionary Computation. Vancouver, USA, 2006: 267-274

[14] Wiratunga N, Lothian R, Massie S. Unsupervised Textual Feature Selection // Proc of the 8th European Conference on Case-Based Reasoning. Fethiye, Turkey, 2006: 340-354

[15] Wiratunga N, Koychev I, Massie S. Feature Selection and Generalisation for Retrieval of Textual Cases // Proc of the 7th European Conference on Case-Based Reasoning. Madrid, Spain, 2004: 806-820

[16] Agrawal R, Imilinski T, Swami A. Mining Association Rules between Sets of Items in Large Database // Proc of the ACM SIGMOD Conference on Management of Data. Washington, USA, 1993: 207-216

[17] University of California-Irvine. UCI Repository of Machine Learning Databases [DB/OL]. [2008-05-06]. http://www.ics.uci.edu/mlearn/-MLRepository.html