ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 6,Nov. 2017,pp. 1232－1238
DOI:10. 16337/j. 1004-9037. 2017. 06. 019

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种改进的基于规则实例多覆盖分类算法

周忠眉[1]　李莎莎[2]

(1. 闽南师范大学计算机学院,漳州,363000; 2. 安徽广播电视大学,合肥,230000)

摘　要: 基于规则分类算法提取的规则集通常存在3个问题:首先,提取的分类规则集中短规则过少,致使高质量的规则不多;其次,规则集中规则数量少,训练数据中几乎所有实例仅被规则覆盖一次;第三,虽然提取大量的规则,但是训练数据中存在一些小类样本的实例不能被任何一条规则覆盖。本文提出一种改进的基于规则的实例多覆盖分类算法(Rule-based classification with instances covered by multiple rules, RCIM),其特点是:(1)为了提高规则的质量,在选择生成规则的第1项时不仅考虑属性值的好坏,而且还考虑了属性值补的好坏;(2)一次产生尽量多,高质量的规则,而且当训练数据的实例至少被两条规则覆盖后才将其删除;(3)当遇上难以判断的测试数据时,对测试数据的各个属性值进行二次学习提取规则。算法RCIM不仅可以有效地提取大量的规则,而且较大程度地提高了规则的质量。通过在大量数据上实验结果表明,RCIM比许多其他算法取得了更高的分类准确率。

关键词: 数据挖掘;分类;分类规则;规则覆盖
中图分类号: TP301　　**文献标志码**:A

Improved Rule Based Classification Algorithm with Multiple Covering Instances

Zhou Zhongmei[1], Li Shasha[2]

(1. College of Computer, Minnan Normal University, Zhangzhou, 363000, China; 2. Auhui Open University, Hefei, 230000, China)

Abstract: There are three problems in rule set which is extracted based on classification algorithm. First, too few short rules in the extracted classification rule set decrease the number of high quality rules. Second, there are such few rules in rule set that almost all of the examples in the training data can be covered only once. Third, despite lots of extracted rules, some examples of small classes in the training data fail to be covered by any of these rules. Herein, a modified example multiple coverage classification algorithm RCIM, which is based on generated rules, is proposed. Here are the features: (1) for the purpose of improving the quality of rules, not only the quality of attribute value but also that of its complement can be taken into account when choosing the first item of a generated rule. (2) It can generate high quality rules at a time as many as possible. (3) It deletes the examples in the training data only if they are covered at least twice. What's more, it can restudy each of the attribute value of the test data to extract rules when encountering the data difficult to judge. The algorithm RCIM not only can efficiently extract a large quantity of rules but also largely improve the quality of rules. Experimental results in many data show that RCIM has achieved higher classification accuracy than many other algorithms.

Key words: data mining; classification; classification rule; rule covering

---

基金项目:国家自然科学基金(61170129)资助项目。
收稿日期:2015-05-29;**修订日期**:2015-06-09

引 言

分类是数据挖掘的一项重要技术,在医疗疾病诊断、图像声音识别和金融电信诈骗等多个方面都有许多应用。分类的目的就是预测未知实例的类别,因此分类好坏的一个重要指标是预测的准确率。基于规则的分类方法指利用训练数据提取一组规则作为分类模型,并基于此模型进行未知实例的预测。基于规则的分类方法如果提取的规则数量少或规则质量不高均不能取得很高的分类准确率。目前基于规则的分类方法主要不足在:(1)提取的规则数量偏少,不能达到训练数据中的一个实例被多条规则覆盖。(2)规则质量不够高。规则质量不高主要是提取的规则长度过长,造成规则的支持度偏低,不易与未知实例匹配。(3)当训练数据不平衡时,训练数据中小类的一些实例甚至未能被任何一条规则覆盖。

选最优值连接规则学习方法(First order inductive learner, FOIL)算法[1]是基于规则的分类方法。首先,它一次选择一个最优属性值,并在其条件数据库中选取最优属性值与其连接产生一条规则;其次,当一条规则生成后移除被此规则覆盖的所有训练实例;最后,循环此步骤,直至训练数据中所有的实例被删除。由于训练数据中的实例只要被规则覆盖一次就被删除,FOIL 算法提取分类规则集的速度较快,但是分类规则集中规则的数量通常很少,导致在不少情况下分类准确率不高。一次选多个优秀值的关联分类算法(Classification based on predictive association rules, CPAR)[2]在分类准确率上优于 FOIL 算法,其方法主要特点是:一次选取 k 个最优属性值,对每个选取的最优属性值都选取其条件数据库中的若干个最优属性值与其连接生成规则。因此,CPAR 算法不仅一次能产生许多分类规则,而且产生的规则集中规则数量比 FOIL 算法多得多,这也是 CPAR 算法在分类准确率上优于 FOIL 算法一个重要原因。决策树算法[3]首先选择一个最优的属性,用此属性的所有属性值作为规则的第 1 项;其次,分别在这些属性值的条件数据库中选择最优属性,并用此属性的每个属性值与其连接生成规则。虽然决策树算法能使得规则集的规则覆盖训练数据中所有的实例,但是提取的规则数量偏少,这些实例仅能被规则覆盖一次,因此,决策树算法在某些数据上的分类准确率同样不高。李莎莎提出改进的基于实例双覆盖决策树算法(Classification based on pruning and double covered rule sets, CDCR-P)算法[4],其主要思路是:首先,选取一组覆盖所有训练数据实例的好的属性值,对每个选取的属性值分别建立条件数据库,并在每个所建立的条件数据库中利用决策树算法提取规则;其次,将训练数据进行分组,并对每组数据二次学习提取规则。用 CDCR-P 算法可以使得训练数据中的每个实例能被规则覆盖两次,实验结果比决策树好,但由于提取的规则数量仍然不够多,分类准确率在一些数据上不如关联分类方法。

通常的关联分类方法主要有两个步骤:(1)给定最小支持度和最小置信度。(2)提取满足所有给定参数的一组关联分类规则进行分类,如关联分类(Classification base of association, CBA)算法[5]和多覆盖关联分类(Classification based on multiple class-association rules, CMAR)算法[6]等。由于关联分类方法提取了非常多的关联分类规则,致使分类模型的复杂度较高,因此其分类准确率也普遍高于一些传统的基于规则的算法,如决策树算法、FOIL 算法。然而,当训练数据是不平衡数据时,由于关联分类中最小支持度的限制,训练数据中样本较少类的一些实例经常不能被任何关联分类规则所覆盖;此外,关联分类规则集中不少规则的置信度偏低,不能有效地预测未知实例。因此,通常的关联分类方法也只是在部分数据上取得了较好的分类准确率。

为了提取更多短的规则,提高规则的支持度,改进分类规则集的质量,从而提高分类的准确率,本文提出一种改进的基于规则实例多覆盖的分类算法(Rule-based classification with instances covered by multiple rules, RCIM)。RCIM 算法的主要有 4 个特点。(1)在提取规则第 1 项时,不仅考虑所有属性值本身的好坏,更重要的是还考虑了属性值补的好坏。从文献[7]可以看出,在某种度量下,属性值的补有时比属性值本身更优。考虑属性值补集的关联分类算法 CCCS[8]提出(Complement class support, CCS)方式来度量属性值的补提取关联分类规则。(2)两两连接提取的满足条件的属性值和属性值的补,并生

成长度为 1 和 2 的短的分类规则。(3)建立候选集和种子集,一次生成尽量多的规则,并确保每个训练数据的实例至少被提取的规则覆盖两次以上。(4)当未知实例无法判断时,采用 Lazy[9] 学习方法。利用该未知实例的属性值,建立新的训练集,进行二次学习提取规则。实验结果表明 RCIM 算法比其他许多算法取得了更高的分类准确率。

1 相关定义

为了较好地度量属性值、属性值的补以及每条规则的好坏,采用支持度、自信度、FOIL 增益和提升度度量属性值的好坏。采用 Laplace 强度度量每条分类规则的好坏。由于支持度和自信度的定义可以参考众多的文献,因此下面仅给出 FOIL 增益、提升度和 Laplace 强度的定义。

定义 1 每个属性值 v 的 FOIL 增益标记为 Gain(v),定义为

$$\text{Gain}(v)=|P^*|(\log|P^*|/(|P^*|+|N^*|)-\log|P|/(|P|+|N|)) \tag{1}$$

式中:$|P^*|$为训练集的所有正例中含有属性值 v 的数目;$|N^*|$为训练集的所有负例中含有属性值 v 的数目;$|P|$为训练集中所有正例的数目;$|N|$为训练集中所有负例的数目。

定义 2 每个属性值 v 的提升度标记为 Lift(v),定义为

$$\text{Lift}(v)=P(v\cup c)/P(v)P(c) \tag{2}$$

式中:$P(v\cup c)$为训练集中的实例,同时含有属性值 v 和类别 c 的概率,$P(v)$和 $P(c)$分别为训练集中属性值 v 出现的概率和类别 c 出现的概率。

Lift(v)>1 表示属性值 v 和类别 c 正相关,Lift(v)<1 表示属性值 v 和类别 c 负相关。Lift(v)=1 表示属性值 v 和类别 c 不相关。

定义 3 (Laplace 强度)每条规则 r 的 Laplace 强度标记为 Laplace(r),定义为

$$\text{Laplace}(r)=(N_c+1)/(N_{\text{tot}}+K) \tag{3}$$

式中:N_c 为规则 r 被训练集类别 c 中实例覆盖的次数,N_{tot}为规则 r 被训练集中实例覆盖的次数,K 为训练集中类别的数目。

由定义 3 可知,即使两条规则的自信度一样,但是这两条规则的 Laplace 强度不一定一样。

2 RCIM 算法与预测

2.1 RCIM 算法提取规则建立分类器

在训练数据中总是假定某个类为正类,其余类的数据实例均看成负类。正类所有规则被提取后,在原始训练集中重新选择另一个类作为正类,同样将其余类的数据实例看成负类,并用同样方法继续提取正类的所有规则,依此类推可以将所有类的规则都提取生成分类规则集。因此,下面例子只给出正类所有规则的提取过程。

例 1 训练数据如表 1 所示,每个对象有 3 个特征属性,最后一列属性 Buy-Computer 为类别属性,有 Yes 和 No 两类,设定训练数据中类别为 Yes 的类作为正类,记为 P,其余类别的数据作为负类,记为 N,提取正类的所有规则。RCIM 算法提取正类所有的规则主要有 4 个步骤,下面分别详述这 4 个步骤:

表 1 训练数据

Tab. 1 Training data

ID	Student	Credit	Income	Buy-Computer
x_1	Yes	Excellent	Medium	No
x_2	Yes	Excellent	High	Yes
x_3	Yes	Excellent	Very high	Yes
x_4	No	Fair	High	No
x_5	No	Excellent	Medium	No
x_6	No	Fair	High	No
x_7	No	Excellent	Very high	Yes
x_8	No	Fair	High	Yes
x_9	No	Excellent	Very high	No
x_{10}	No	Fair	High	No

(1)提取满足条件的属性值和属性值的补,生成长度为 1 和 2 的短的分类规则。

设定最小支持度和最小自信度均为 30%,计算训练集中每个属性值和属性值补的 FOIL 增益和提升度,表 2 是类别为 Yes 实例的所有属性值和属性值补的 FOIL 增益 Gain 和提升度 Lift。设 F_1 为表 2 中所有满足给定的最小支持度和最小自信度且提升度大于 1 的属性值和属性值的补的集合,则 F_1 为 F_1={Student=Yes,Credit=Excellent,Income=Veryhigh,Income≠Medium},将 F_1 中的元素两两连接生成长度为 2 的项集,设 F_2 为 F_1 中元素两两连接后满足最小支持度和最小自信度,且提升度大于 1、长度为 2 的项集集合,若 F_1 和 F_2 中元素的自信度为 100%,则直接生成规则,并加入分类规则集中。经过计算,F_2 中的两个元素 Student=Yes∧Income=Veryhigh 和 Student=Yes∧Income≠Medium 的自信度为 100%,直接生成类别为 Yes 的规则,F_2 中其余元素为 Student=Yes∧Credit=Excellent, Credit=Excellent∧Income=Veryhigh 和 Credit=Excellent∧Income≠Medium。

表 2 属性值以及属性值补的 Gain 和 Lift 值

Tab. 2 Gain and lift values of attribute-value pairs and their complements

属性值	Gain	Lift	属性值的补	Gain	Lift
Student=Yes	0.443 7	>1	Student=No	−0.292 3	<1
Credit=Excellent	0.290 7	>1	Credit=Fair	−0.204 1	<1
Income=High	0.000 0	=1	Income≠High	0.000 0	=1
Income=Very high	0.443 7	>1	Income≠Very high	−0.292 3	<1
Income=Medium	0	<1	Income≠Medium	0.387 6	>1

(2)建立候选集和种子集,生成规则。

将集合 F_1 和集合 F_2 合并建立候选集,并选择 F_1 中 FOIL 增益值 Gain>0 的元素作为种子集,并将种子集中元素按照 FOIL 增益值由大至小进行排序。对候选集中每个元素,RCIM 算法生成两个模式。设 x 为候选集中的任意一个元素,选取 x 的条件数据库中 FOIL 增益最大的属性值与 x 连接生成第 1 个模式 X_1,属性值 x 的条件数据库由训练数据中含有 x 的所有实例组成。另外,选取 x 的条件数据库中存在的,种子集中 FOIL 增益值最大的元素与 x 连接生成第 2 个模式 X_2。

表 3 是候选集中元素 Credit=Excellent 的条件数据库的所有属性值的 FOIL 增益 Gain 的值。从表 3 可以看出,在 Credit=Excellent 的条件数据库中具有最大 FOIL 增益值的元素为 Student=Yes,在 Credit=Excellent 的条件数据库中存在的,种子集中 FOIL 增益值最大的元素为 Income=Veryhigh,将这两个元素分别与元素 Credit=Excellent 连接,得到的两个模式为 Credit=Excellent∧Student=Yes 和 Credit=Excellent∧Income=Veryhigh.

如果这些模式的置信度为 100%,则直接将这些模式生成规则,反之,如果这些模式的置信度不足 100%,而这些模式的置信度比连接前的置信度高,则继续选取这些模式条件库数据中最好的属性值与其连接生成规则。反之,如果这些模式的置信度比连接前的置信度低,则这些模式停止生成规则。但这些模式的置信度如果足够高,则将其作为备用规则。在测试时,备用规则与二次学习提取的规则集合并来预测未知实例的类别。提取规则的过程见算法 1。

表 3 Credit=excellent 条件库中所有属性值的 Gain 值

Tab. 3 Gain values of attribute-value pairs in Credit=excellent conditional database

属性值	Gain
Student=Yes	0.249 9
Student=No	−0.176 1
Income=High	−0.176 1
Income=Very high	0.249 9
Income=Normal	0.000 0

(3)如果正类的实例被规则覆盖两次及以上,则删

除正类的实例,完成一次正类规则的提取。

RCIM 算法将一个候选集的所有元素都生成规则后,对提取的规则集,检查正类中实例被提取规则的覆盖情况。如果正类中实例被规则覆盖两次及以上,则此实例被删除。如果正类中不存在实例被提取的规则覆盖两次以上,则降低设定的最小支持度和最小置信度的阈值,重新建立候选集,提取规则集。

(4)删除所有正类的实例,完成正类所有规则的提取。

每次正类的实例被删除后,若正类中还有实例存在,则重复规则的提取过程,循环反复直至正类的实例被全部删除。规则提取过程见算法 2。

算法 1 RCIM 算法根据候选集生成规则

输入:训练集 T,候选集 F_1, F_2 输出:规则集 R

规则集 R 为空,初始化队列 Q

Q. push(F_1), Q. push(F_2)

while ! Q. empty()

pattern x = Q. front(); Q. pop()

if x 条件库不为空

计算条件库中的最优值 p_1,选择种子集最优值 p_2

连接 x 与 p_1, p_2

if(连接后($x+p_i$). confidence==100)//置信度 100%时

$R=R\{x+p_i\}$

else if(连接后($x+p_i$). confidence x. confidence) //置信度有提升

Q. push()

else continue;//置信度无提升

end if

end if

end while

算法 2 RCIM 算法提取规则

输入:训练集 T,最小支持度 $\text{sup}_{\min}$,最小置信度 $\text{confi}_{\min}$

输出:分类规则 R

规则集 R 为空; 候选集 F_1, F_2 为空

while $|p|>0$

计算属性值 v_i 以及 v_i 的支持度 support,置信度 confidence,相关度 lift

满足给定最小支持度 $\text{sup}_{\min}$,最小置信度 $\text{confi}_{\min}$ 及 lift >1 等条件的属性值得到规则和关联 1-项集 F_1

由关联 1-项集 F_1 得到关联 2-项集 F_2,建立候选集 candidate

for each item p. incandidate

找出属于 p 的规则集 r

$R=R\cup\{r\}$

end for

若一条实例被规则覆盖两次及以上,删除被规则覆盖的实例

end while

2.2 预测新实例

基于规则的分类算法通常提取大量置信度为100%的规则,为了区分这些规则的质量,使用Laplace强度度量规则的好坏,并依此度量由大到小规则的排序。对每个待测的实例,找出规则集中与此实例匹配的所有规则,并对每个类别选出前3条规则。依次计算各类别中规则的平均Laplace强度,并选取平均强度最大的类别作为此待测实例的类别。若待测实例按此方法无法判断其类别时,利用Lazy的学习方法进行二次学习。从训练集中选取含有待测实例任意一个属性值的样本构建新训练集,进行二次学习,提取另外的规则集,并将此规则集与备用规则合并,对合并后规则集中的规则按照置信度和支持度的大小由大到小排序,选取最前面的规则进行预测。

3 RCIM算法实验结果与分析

RCIM算法在20个UCI数据集上与3种经典算法CBA,CMAR及CPAR进行实验结果对比,实验与测试采用10-折交叉验证方法,即将数据集10折,依次取其中的9折作为训练集,剩余的1折作为测试集,实验结果为10次测试结果的平均值。实验中设置最小支持度与最小置信度均为10%,给定FOIL增益Gain的最小阀值为0.5。在规则提取的过程中当模式的置信度大于等于60%时,保留为备用规则。

表4和图1分别给出了上述4种算法CBA,CMAR,CPAR和RCIM的分类准确率。表4的最后一行给出了每个算法在20个数据集上的平均分类准确率。由表4可以得出RCIM算法的平均分类准确率均高于其他3种算法。图1更直观地表明4种算法在各个数据集的分类准确率,图1同样显示了RCIM算法取得了最好的实验结果。

表4 算法RCIM与CBA,CMAR,CPAR准确率的对比

Tab.4 Accuracy of CBA,CMAR,CPAR and RCIM

数据集	属性	类别	实例	CBA	CMAR	CPAR	RCIM
Austral	14	2	690	0.849	0.861	0.862	0.862 319
Auto	25	7	205	0.783	0.781	0.820	0.820 000
Breast	10	2	699	0.963	0.964	0.960	0.964 203
Cleve	13	2	303	0.828	0.822	0.815	0.838 280
Diabetes	8	2	768	0.745	0.758	0.751	0.781 200
German	20	2	1 000	0.734	0.749	0.734	0.751 000
Glass	9	7	214	0.739	0.701	0.744	0.673 160
Heart	13	2	270	0.819	0.822	0.826	0.840 700
Hepatic	19	2	155	0.818	0.805	0.794	0.865 417
Horse	22	2	368	0.821	0.826	0.842	0.829 100
Iono	34	2	351	0.923	0.915	0.926	0.931 825
Iris	4	3	150	0.947	0.94	0.947	0.933 300
Labor	16	2	57	0.863	0.897	0.847	0.863 333
Lymph	18	4	148	0.778	0.831	0.823	0.831 429
Pima	8	2	768	0.729	0.751	0.738	0.781 200
Sonar	60	2	208	0.775	0.794	0.793	0.802 857
Tic-tac	9	2	958	0.996	0.992	0.986	0.984 300
Vehicle	18	4	846	0.687	0.688	0.695	0.699 874
Wine	13	3	178	0.950	0.950	0.955	0.988 900
Zoo	16	7	101	0.968	0.971	0.951	0.960 900
平均				0.842 6	0.847 2	0.847 3	0.850 400

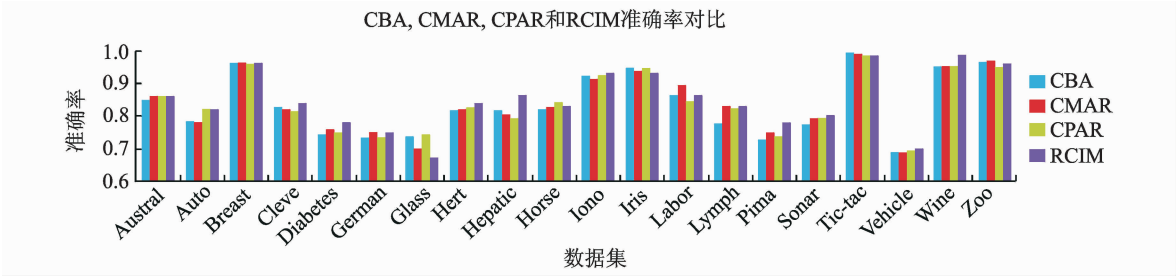


图 1 CBA,CMAR,CPAR 和 RCIM 的准确率对比

Fig. 1 Accuracy comparison of CBA,CMAR,CPAR and RCIM

4 结束语

对于基于规则的分类模型,分类规则的质量和分类规则的数量是影响分类准确率的两个重要因素。在分类模型中,如果一些规则的质量不高,将大大影响对未知对象类别的准确判断;同时,如果规则的数量不够,则大量有用的信息未能被提取,也将导致对未知对象类别的错判。本文提出的基于实例多覆盖的规则提取算法使训练集的每个实例至少被规则覆盖两次,表明分类模型中有足够数量的规则。其次,该算法尽量产生更多短的规则,以提高规则的质量。在 20 个数据集上与其他经典算法对比的实验结果表明,本文算法取得了较高的分类准确率。

参考文献:

[1] Ross Q J, Cameron-Jones R M. FOIL: A midtern report[C]// Machine Learning: ECML-93, European Conference on Machine Learning. Vienna, Austria: [s. n.], 1993:3-20.

[2] Yin Xiaoxin, Han Jiawei. CPAR: Classification based on predictive association rules[C]//The 2003 SIAM International Conference on Data Mining. California, USA: [s. n.], 2003,5.

[3] Ross Q J. Induction on decision trees[J]. Machine Learning, 1986,1(1):81-106.

[4] Li Shasha, Zhou Zhongmei, Wang Weiping. Classification based on pruning and double covered rule sets for the internet of things applications[J]. Scientific World Journal, 2014,1(1):1-6.

[5] Liu B, Hsu W, Ma Y. Integrating classification and association rule mining[C]// Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98). New York, USA: [s. n.], 1998:80-86.

[6] Li Wenmin, Han Jiawei, Pei Jian. CMAR: Accurate and efficent classification based on multiple class-assocation rules[C]// ICDM'01. San Jose, CA:[s. n.], 2001:369-376.

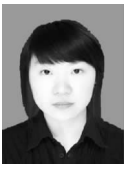
[7] An Aijun. Learning classification rules from data[J]. International Journal of Computers and Mathematics with Applications, 2003,4(4/5):737-748.

[8] Bavani A, Sanjay C. CCCS: A top-down associative classifier for imbalanced class distribution[C]// KDD'06 Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA:[s. n.], 2006: 517-522.

[9] Veloso A, Jr Meira W, Zaki M J. Lazy associative classification[C]//ICDM 06. Hong Kong, China:[s. n.], 2006:645-654.

作者简介:

周忠眉(1965-),女,博士,教授,研究方向:数据挖掘、人工智能,E-mail: zzm@zju.edu.cn。



李莎莎(1988-),女,硕士研究生,研究方向:数据挖掘。