基于相关规则的不平衡数据的关联分类

黄再祥 周忠眉 何田中
(漳州师范学院计算机科学与工程系 漳州 363000)

摘 要 许多研究表明关联分类具有较高的分类准确率,然而,大多数关联分类基于"支持度-置信度"框架,在不平衡数据集中,置信度和支持度都偏向产生多数类的规则,因此,少数类的实例容易被错误分类。针对上述问题,提出了一种基于相关规则的不平衡数据的关联分类算法。该算法挖掘频繁且互关联的项集,在以该项集为前件的分类规则中选取提升度最大的规则。规则按结合了提升度、置信度和补类支持度(CCS)的规则强度进行排序。实验表明,该算法取得了较高的平均分类准确率且在分类少数类的实例时具有更高的准确率。

关键词 数据挖掘,关联分类,不平衡数据,相关规则
中图法分类号 TP311.13 **文献标识码** A

Correlated Rules Based Associative Classification for Imbalanced Datasets

HUANG Zai-xiang ZHOU Zhong-mei HE Tian-zhong
(Department of Computer Science and Engineering, Zhangzhou Normal University, Zhangzhou 363000, China)

Abstract Many studies have shown that associative classification is a promising classification method. However, most algorithms of associative classifications may not achieve high classification performance on imbalanced datasets because they generate rules based on the "support-confidence" framework. The confidence (support) tends to bias the majority class in imbalanced datasets. As a result, these instances with minority class may be misclassified. We proposed a new associative classification approach called CRAC (Correlated Rules based Associative Classification for Imbalanced Datasets). First, we mine frequent and mutual associative itemsets for classification. Therefore, we will generate small set of high-quality rules. Second, CRAC only select the rule with largest lift as a CAR among all rules with that frequent and associative itemset as condition. As a result, the antecedent and the consequent of the rules CRAC generated are positively correlated. Finally, we rank rules according to a new metric which integrates lift, support and Complement Class Support (CCS). So, we are likely to use rules with positively correlation to prediction the minority class. Our experiments on fifteen UCI data sets show that our approach is an effective classification technique for both balance and imbalanced datasets, and has better average classification accuracy in comparison with CBA.

Keywords Data mining, Associative classification, Imbalance datasets, Correlated rules

1 引言

作为数据挖掘的主要任务之一,分类得到了广泛研究。分类的目标是根据训练数据建立分类器来预测类标未知的实例。1998年,Liu等提出一种被称为关联分类的新分类方法[1]。关联分类将关联规则挖掘集成到分类中。此后,提出了一批准确高效的关联分类算法,比如,CMAR[2]、CPAR[3]、CAEP[4]、HARMONY[5]。这些算法显示关联分类算法的准确率比决策树方法如C4.5[6]要高。

在许多分类的应用领域,比如医疗诊断、欺骗侦测,数据集中的类分布是不平衡的,即一个类的实例比其他类的实例少得多。然而,由于大多数关联分类算法采用"支持度-置信度"框架产生规则,使得少数类的实例较难分类准确。主要有两个原因:(1)在不平衡数据集中少数类的支持度很低。如果支持度阈值设置较高,很难挖掘到属性值与少数类之间的关联。另一方面,如果为得到少数类的规则将支持度阈值设置较低,将产生大量的多数类的规则,其中包含了大量冗余或噪声。(2)置信度也会偏向多数类[7]。在不平衡数据集中,置信度高的规则可能是负相关规则。例如,数据集中有两个类,支持度分别为 $\sup(C_1)=90\%$,$\sup(C_2)=10\%$。由频繁集 X 产生两条规则:

规则1 $X\rightarrow C_1$,置信度70%

规则2 $X\rightarrow C_2$,置信度30%

如果按置信度高来选择规则,规则1将作为分类规则。然而规则1的前件与后件是负相关,而正相关的规则2被丢弃了。因此,选择置信度高的规则将很可能丢失少数类的正相关规则。另外,如果按置信度和支持度降序排列规则,那么少数类的规则的优先级比较低,从而有可能使用多数类的规

到稿日期:2013-05-20 返修日期:2013-08-09 本文受国家自然科学基金(61170129),福建省自然科学基金(2013J01259)资助。
黄再祥(1975—),男,讲师,CCF会员,主要研究方向为数据挖掘,E-mail:huangzaixiang@126.com;周忠眉(1965—),女,博士,教授,主要研究方向为数据挖掘、人工智能;何田中(1970—),男,讲师,CCF会员,主要研究方向为数据挖掘。

则来预测少数类的实例。

本文提出了一种基于相关规则的不平衡数据的关联分类方法。该方法与其他关联分类方法主要有以下 3 个方面不同：

(1)挖掘频繁互关联的项集生成规则。我们采用 all-confidence 度量项集中项之间的关联程度。

(2)在以频繁互关联项集为前件的所有规则中选择提升度最大的规则，丢弃其他的规则。提升度能保证挖掘到的规则具有正相关性。

(3)提出了一种新的度量来排序规则。该度量集成了提升度、支持度和 CCS[8]。

2 基本概念及相关工作

2.1 基本概念

关联分类方法是一种使用一组关联规则来分类事务的技术。它挖掘形如 $X\Rightarrow C_i$ 的类关联规则，其中 X 是项(或属性-值对)的集合，而 C_i 是类标号。假设有两个规则 $R_1:X_1\Rightarrow C_1$ 和 $R_2:X_2\Rightarrow C_2$；如果 $X_1\subset X_2$，则称 R_1 是 R_2 的泛化规则，R_2 是 R_1 的特化规则。

在关联分类中，设训练集 $T=\{t_1,t_2,\cdots,t_n\}$ 有 m 个不同的特征属性 $A_1,A_2,\cdots,A_m$ 和一个类属性 C。如果实例 $t_i\supseteq X$，称之为实例 t_i 的匹配规则 $X\Rightarrow Y_i$。

支持度和置信度是关联规则的两个重要度量。支持度(s)确定项集 X 可以用于给定数据集的频繁程度，而置信度(c)确定 C_i 在包含 X 的事务中出现的频繁程度。这两种度量的形式定义如下：

$$s(X)=|\{t_i\,|\,X\subseteq t_i,t_i\in T\}|/|T| \tag{1}$$

$$c(X\Rightarrow C_i)=s(X\bigcup C_i)/s(X) \tag{2}$$

如果对于规则 $R:X\Rightarrow C_i$，有 $c(X\Rightarrow C_i)>s(C_i)$，这样的规则称为正相关规则。

Omiecinski 提出了 all-confidence 的概念[9]。项集 $X=(a_1,\cdots,a_n)$ 的 all-confidence 定义如下：

$$allconf(X)=s(X)/\max(s(a_1),\cdots,s(a_n)) \tag{3}$$

all-confidence 代表从一个项集中抽出的所有关联规则中的最小置信度。all-confidence 是项集中项之间关联程度的一种很好的度量。如果项集的 all-confidence 超过用户指定的阈值，则称该项集为互关联项集。all-confidence 也可以用来定义规则中项集与类的关联程度，定义如下：

$$allconf(X\Rightarrow Y_i)=s(X\bigcup Y_i)/\max(s(X),s(Y_i)) \tag{4}$$

信息熵[10]是信息量的一种度量。项集 X 的信息熵 $E(X)$ 定义如下：

$$E(X)=-\frac{1}{\log_2 k}\sum_{i=1}^{k}p(Y_i\,|\,X)\log_2 p(Y_i\,|\,X) \tag{5}$$

式中，k 是类标的个数，$p(Y_i\,|\,X)$ 是匹配 X 的实例属于 Y_i 的概率。项集信息熵越大，分类的信息量越小。因此，可以提前删除信息熵接近 1 的项，从而提高挖掘类关联规则的效率。

2.2 相关工作

Liu 等于 1998 年提出了第一个关联分类算法 CBA。CBA 采用 Apriori 算法[11]来挖掘频繁项集。它需要多趟扫描数据库来计算支持度。CBA 将置信度超过阈值的类关联规则作为候选规则，并按置信度、支持度将候选规则排序，根据数据库覆盖原理选择一个规则子集构建分类器。

为了加快频繁项集和规则产生的效率，MMAC[12]采用了一种基于交集方法的新技术。MMAC 扫描训练集一次记录单个项出现的行号。通过两个频繁 $k-1$ 项集的行号的交集很容易得到频繁 k 项集的支持度。

Arunasalam 等提出了一种称为 CCS(Complement Class Support)的度量，定义如下[8]：

$$CCS(X\rightarrow C_i)=\sup(X\overline{C_i})/\sup(\overline{C_i}) \tag{6}$$

CCS 捕获了规则在补类中的强度。一个强规则的 CCS 应该较小。

3 基于相关规则的关联分类

本文提出了一种新的关联分类算法，称其为基于相关规则的不平衡数据的关联分类(CRAC)。它包含 3 个阶段：(1)规则产生，主要挖掘频繁且互关联的项集产生规则；(2)规则排序，排序时综合考虑规则的提升度、支持度和 CCS；(3)规则剪枝，采用了 3 种剪枝技术减少冗余规则。

3.1 规则产生

CRAC 挖掘频繁且互关联的项集产生规则。所谓频繁且互关联项集是支持度、all-confidence 都超过相应阈值的项集。一旦找到频繁互关联项集，就可以产生所有以该项集为前件的规则，我们只选择具有最大提升度的规则。

3.2 规则排序

为了选择合适的规则来分类新实例，大多数关联分类通常将规则进行排序。关联分类中规则排序起到很重要的作用。CBA 主要根据置信度和支持度排序规则。当几个规则有相同的置信度和支持度时，CBA 随机选择其中的一个规则，这样可能降低分类的准确率。这种方法偏好高置信度的规则，然而，使用最高置信度规则来分类并不总是正确的。例如，假设我们要确定顾客的信用限额。数据集有 3 个属性(no_job，investment_immigrant，oversee_asset>500k)。假设匹配一个顾客的最高置信度的两个规则如下：

规则 1　no_job→credit_limit_3000－(支持度：3000，置信度：95%)

规则 2　investment_immigrant→credit_limit_3000＋(支持度：5000，置信度：94%)。

根据置信度排序，我们将根据规则 1 将该顾客分类到 credit_limit_3000－。然而，两个规则的置信度相差很小，而规则 2 具有更强的支持度。

为了解决这个问题，我们按规则的强度来排序规则。规则强度集成了提升度、支持度和 CCS，定义如下：

$$SR(X\rightarrow C_i)=\frac{lift(X\rightarrow C_i)*\sup(XC_i)}{\max(CCS(X\rightarrow C_i),t)} \tag{7}$$

当一个规则的置信度为 100%时，其 $CCS(X\rightarrow C_i)=0$，为了避免除以 0，我们给 t 设置一个很小的值如 0.001。

规则按上述强度排序将增加少数类的规则预测新实例的机会。

3.3 规则剪枝

关联分类方法通常产生大量的规则。为了提高分类的效率，需要剪去冗余规则。我们的算法采用以下 3 种方法对规则进行剪枝。

(1)删除具有高信息熵的单个项。当信息熵接近 1 时，这样的数据项具有很少的分类信息。

(2)利用泛化规则剪枝特殊规则。假设有两个规则 $R1$ 和 $R2$，$R1$ 是 $R2$ 的泛化规则。如果 $R1$ 的置信度大于 $R2$ 的置信度，剪去规则 $R2$。

(3)与 CBA 类似，根据数据库覆盖[1]剪枝规则。

3.4 CRAC 算法

CRAC 算法的细节如算法 1 所示。

算法 1 CRAC

输入：数据集 D；支持度阈值 minsup；all-confidence 阈值 minallconf；信息熵阈值 maxentropy。

输出：规则集 R。

1. C_1←init-pass(D)；
2. ruleGen(C_1，L_1，R)；
3. for(k=2；$L_{k-1}\neq\Phi$；k++) do
4. C_k←candidateGen(L_{k-1})；
5. ruleGen(C_k，L_k，R)；
6. end for
7. Sort(R)；
8. DatabaseCoverage(R，k)；
9. return R；

算法中，C_k 表示候选 k 项集，L_k 是频繁且互关联 k 项集，R 为规则集。在第一趟扫描数据集中(行 1)，记录每一个项的类分布行号。然后执行 ruleGen 进行规则抽取(行 2)。在接下来的循环中，主要执行两个主要的操作：候选 k 项集的产生(行 4)和规则产生(行 5)。当规则产生后，根据规则强度对规则进行排序(行 7)，然后使用数据库覆盖方法进行剪枝(行 8)得到最终规则集。

candidateGen 方法与 Apriori 算法中的 apriori-gen 方法类似，利用($k-1$)项集进行连接运算得到候选 k 项集。不同的是，candidateGen 同时对两个 $k-1$ 项集的类分布行号进行交运算，得到 k 项集的类分布行号，便可得项集在每个类中的支持度。例如，假设某数据集有两个类 C_1 和 C_2，有两个项集的类分布如下：

$\langle X_1,\{(C_1,\{1,2,3\}),(C_2,\{6,7,8,9\})\}\rangle$

$\langle X_2,\{(C_1,\{2,3,4\}),(C_2,\{5,6,7,8,10\})\}\rangle$

通过求交运算可得$\langle X_1\cup X_2,\{(C_1,\{2,3\}),(C_2,\{6,7,8\})\}\rangle$。

方法 ruleGen 的细节如算法 2 所示。

算法 2 ruleGen(C_k，L_k，R)

1. for all X∈C_k do
2. 计算 X 的 all-confidence；
3. if(s(X)≥minsup && allconf(X) ≥minallconf)
4. 计算以 X 为前件的所有规则的提升度并选择具有最大提升度的规则 R_i：X→C_i
5. 在 R 中找出 R_i 的所有泛化规则
6. if R_i 的所有泛化规则的置信度都小于 R_i 的置信度
7. 将 R_i 加入 R 中；
8. end if
9. 将 X 加入 L_k中；
10. end if
11. end for

挑出支持度和 all-confidence 都超过阈值的项集，计算该项集到每个类的提升度，只选具有最大提升度的规则(行 3—4)。如果 R_i 的所有泛化规则的置信度都小于 R_i 的置信度，则将 R_i 作为候选规则(行 5—8)。

4 实验结果及分析

我们进行了广泛的实验来评估 CRAC 算法的分类准确率。所有的实验在 3GHz、Pentium 4、1G 内存的 PC 机上进行。我们采用 UCI 机器学习库中的 13 个数据集进行 10 折交叉验证实验。13 个数据集的特征如表 1 所列。

表 1 UCI 数据集特征

数据集	属性数	类别数	实例数
Anneal	38	6	898
Austral	14	2	690
auto	25	7	205
Breast	10	2	699
Cleve	13	2	303
Crx	15	2	690
Diabetes	8	2	768
Heart	13	2	270
Horse	22	2	368
Hypo	25	2	3163
Iris	4	3	150
Lymph	18	4	148
Zoo	16	7	101

表 2 显示了 CBA 和 CRAC 两种算法在 13 个数据集上的准确率。CBA 算法使用文献[1]的实现程序[13]。为了比较提升度的作用，我们实现了两个版本的 CRAC：CRAC-conf 和 CRAC-lift。CRAC-conf 表示找到一个频繁且互关联的项集后选择具有最大置信度的规则，CRAC-lift 表示找到一个频繁且互关联的项集后选择具有最大提升度的规则。所有程序的支持度阈值都设为 0.5%。CBA 的置信度阈值设为 50%。CRAC 的 all-confidence 的阈值设为 0.1，信息熵的阈值设为 0.95。对于有连续型属性的数据集，采用 CBA 实现算法中提供的方法进行离散化。

结果显示 CRAC-conf 和 CRAC-lift 的平均准确率都比 CBA 高，而 CRAC-lift 具有最高的平均准确率。与 CRAC-conf 相比，CRAC-lift 在 13 个数据集上的准确率都不低于 CRAC-conf，且有 5 个数据集上的准确率有所提高。实验数据说明以提升度选择规则比以置信度选择规则能取得更高的准确率。

表 2 CBA 和 CRAC 的准确率

数据集	CBA	CRAC-conf	CRAC-lift
Anneal	97.1	96.5	**97.9**
Austral	**86.2**	**86.2**	**86.2**
auto	76.1	78.0	**80.0**
Breast	**96.1**	95.9	95.9
Cleve	81.1	**84.7**	**84.7**
Crx	84.3	**85.7**	**85.7**
Diabetes	**75.3**	73.8	74.5
Heart	82.6	**85.2**	**85.2**
Horse	81.7	**85.3**	**85.3**
Hypo	**98.7**	97.8	97.9
Iris	93.9	**94.0**	**94.0**
Lymph	82.5	**84.3**	**84.3**
Zoo	**97.1**	90.0	95.0
平均值	87.1	87.5	**88.2**

我们以 Anneal 数据集为例更深入地考察 CRAC-conf 和 CRAC-lift 的差别。取 10 折交叉验证中的一折数据列于表 3 中，该表显示了两种算法产生的不同类别的规则数对比。其中第 1 行显示了 Anneal 数据集中各类别实例的支持度。从

(下转第 122 页)

[5] Li Li-yuan, Huang Wei-min, Irene Y H, et al. Foreground Object Detection from Videos Containing Complex Background [C]// Proceedings of the eleventh ACM international conference on Multimedia. 2003:2-10

[6] 张文涛,李晓峰,李在铭. 高速密集视频目标场景下的运动分析[J]. 电子学报,2000,28(10):114-117

[7] 杨广林,孔令富. 基于图像分块的背景模型构建方法[J]. 机器人,2007,29(1):29-34

[8] 姚春莲,周兵. 运动对象检测及其在视频压缩与处理中的应用[M]. 北京:冶金工业出版社,2010

[9] 李庆武,王敏,陈琦,等. 基于图像分块和 Hausdorff 距离的背景更新方法[J]. 测试技术学报,2011,25(6):544-547

[10] Jodoin J P, Bilodeau G A, Saunier N. Background subtraction based on local shape [J]. arXiv preprint arXiv:1204.6326,2012

[11] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

[12] Wang Zhou, Bovik A C, Simoncelli. Structural approaches to image quality assessment[J]. Handbook of Image and Video Processing, 2005, 7: 18

[13] Loza A, Mihaylova L, Canagarajah N, et al. Structural Similarity-Based Object Tracking in Video Sequences[C]// 2006 9th International Conference on Information Fusion. IEEE, 2006: 1-6

[14] Łoza A, Mihaylova L, Bull D, et al. Structural Similarity-Based Object Tracking in Multimodality Surveillance Videos [J]. Machine Vision and Applications, 2009, 20(2): 71-83

[15] PETS 2006 benchmark data[OL]. http://www.pets2006.net, 2006

(上接第 113 页)

表 3 中可以看出,CRAC-lift 在少数类 C_2、C_4、C_5 中产生更多的规则数。

表 3 产生各类别的规则数对比

类	C_1	C_2	C_3	C_4	C_5
支持度	76.1%	11.0%	7.5%	4.5%	0.9%
CRAC-conf	21	7	1	1	0
CRAC-lift	18	11	1	2	3

表 4 和表 5 显示了 CRAC-lift 和 CRAC-conf 的混淆矩阵。在 89 个测试实例中,CRAC-lift 将少数类 C_5 的一个实例做出了正确预测,而 CRAC-conf 没有匹配该实例的规则。CRAC-lift 正确分类了 C_2 中的 10 个实例,而 CRAC-conf 只正确分类了 7 个实例,另外 3 个没有匹配规则。因此,CRAC-lift 主要提高了少数类的分类准确率。

表 4 CRAC-lift 在 Anneal 上的混淆矩阵

	C_1	C_2	C_3	C_4	C_5	NoRules
C_1	68			1		
C_2		10				
C_3			6			
C_4				3		
C_5					1	

表 5 CRAC-conf 在 Anneal 上的混淆矩阵

	C_1	C_2	C_3	C_4	C_5	NoRules
C_1	68					1
C_2		7				3
C_3			6			
C_4				3		
C_5					0	1

结束语 在不平衡数据中,正确分类少数类的实例具有更重要的意义。而置信度、支持度等度量都偏向产生更多的多数类的规则,使得少数类的规则较少,甚至没有。本文提出了一种基于相关规则的不平衡数据的关联分类。以提升度来挑选规则,将产生更多少数类的规则。以结合提升度、支持度和 CCS 的强度排序规则,使得少数类的规则的优先级较高,因而在分类少数类实例时,有可能优先选中少数类规则做出正确分类。实验表明,该算法取得了较高的平均分类准确率且在分类少数类的实例时具有更高的准确率。

参 考 文 献

[1] Liu B, Hsu W, Ma Y. Integrating classification and association rule mining[C]// Proc of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98). 1998:80-86

[2] Li W, Han J, Pei J. CMAR: Accurate and efficient classification based on multiple class-association rules[C]// Proc of the 1st International Conference on Data Mining. 2001:369-376

[3] Yin X, Han J. CPAR: classification based on predictive association rules[C]// Proc of the SIAM International Conference on Data Mining (SDM'03). 2003:331-335

[4] Dong G, Zhang X, Wong L, et al. CAEP: Classification by aggregating emerging patterns[C]// Discovery Science. Springer Berlin Heidelberg, 1999:30-42

[5] Wang J, Karypis G. HARMONY: Efficiently mining the best rules for classification[C]// Proc. of SDM. 2005:205-216

[6] Quinlan J R. C4.5: programs for machine learning[M]. Morgan kaufmann, 1993

[7] Verhein F, Chawla S. Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets[C]// Seventh IEEE International Conference on Data Mining, 2007, ICDM 2007. IEEE, 2007:679-684

[8] Arunasalam B, Chawla S. CCCS: a top-down associative classifier for imbalanced class distribution[C]// Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006:517-522

[9] Omiecinski E R. Alternative interest measures for mining associations in databases [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(1):57-69

[10] Zhao Y, Karypis G. Criterion functions for document clustering: Experiments and analysis [Z]. Machine Learning, 2001

[11] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]// Proc of the 20th International Conference on Very Large Data Bases (VLDB'94). 1994:487-499

[12] Thabtah F A, Cowling P, Peng Y. MMAC: A New Multi-class, Multi-label Associative Classification Approach[C]// Proc of the 4th International Conference on Data Mining (ICDM'04). 2004: 217-224

[13] CBA: http://www.comp.nus.edu.sg/dm2/p-download.html