第36卷 第20期 计 算 机 工 程 2010年10月
Vol.36 No.20 Computer Engineering October 2010
·软件技术与数据库· 文章编号：1000—3428(2010)20—0052—03 文献标识码：A 中图分类号：TP311

适用于非平衡数据的多关系多分类模型

杨鹤标，王 健
(江苏大学计算机科学与通信工程学院，江苏 镇江 212013)

摘 要：针对多关系多分类的非平衡数据，提出一种分类模型。在预处理阶段，建立目标类纠错输出编码(ECOC)、目标关系与背景关系间的虚拟连接并完成属性聚集处理，进而划分训练集和验证集。在训练阶段，依据一对多划分思想，结合CrossMine算法构造多个子分类器，采用AUC法评估验证各子分类器。在验证阶段，比较目标类ECOC与各子分类器分类结果连接字的海明距离，选择最小海明距离的目标类为最终分类。经合成和真实数据的实验，验证了模型有效性及分类效果。

关键词：多关系分类；非平衡数据；多类分类；纠错输出编码；一对多划分

Multi-relational Multi-class Model for Imbalanced Data

YANG He-biao, WANG Jian
(School of Computer Sciencre and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China)

【Abstract】 This paper proposes a multi-relational model which is applied to the multi-class imbalanced data. In the preprocessing stage, each class is assigned an Error Correcting Output Coding(ECOC). After setting up the virtual joins between the target and background relations, appropriate aggregation functions are used for different features. On this condition, the data can be divided into training set and validation set. Sub-classifiers are built on the training set in combination with One-vs-All classification method and CrossMine algorithm, and all the sub-classifiers are validated by their AUC values. The ECOC of the target class is compared with the Hamming distance of the linked word produced by the sub-classifiers on the validation set, and the class is chosen which has the shortest Hamming distance for the final result. The validity and effectiveness of the classifier by experiments are shown on both synthetic and real datasets.

【Key words】 multi-relational classification; imbalanced data; multi-class classification; Error Correcting Output Coding(ECOC); One-vs-All classification

1 概述

多关系分类可直接从多关系数据集中寻找有效分类模式，无需将关系数据库中的多个表转换为单一数据表，从而有效避免了传统分类方法中出现的信息丢失、统计偏斜和效率降低等问题。借鉴归纳逻辑程序设计技术，已经形成了多种多关系分类方法，代表性的如CrossMine[1]等。这些方法对两分类的平衡数据集具有较高的分类准确率及良好的性能，然而面对实际应用中多分类的非平衡数据(如在医疗检查、产品质量检测、网络入侵检测等领域)，则无法达到期望的分类效果。在传统分类方法中，业界通过结合已有多分类方法和非平衡数据下两分类策略来解决非平衡数据下的多分类问题，其中包括代价敏感方法、集成学习方法(如构造SVM多类分类器)等。以上方法中代价敏感的误分类代价难以预知；基于SVM的多类分类器，因分类面太多或优化方程太复杂而丧失可操作性，尤其在使用大型数据元组集进行训练时，SVM处理速度很慢，缺少实用性。另外，这些方法只可用于传统单一数据表，无法直接运用于多关系数据。

本文提出了一种适用于非平衡数据的多关系多分类模型。模型经预处理、训练及分类检验完成构建。将其分别运用于合成和真实关系数据集，并验证了其分类效果。

2 相关定义

定义1 关系数据库R中的关系记为：$R=\{T_i\}_1^n$，表T_i的主键记为$T_{i.\text{key}}$，R中的目标关系记为T_{target}，其他背景关系为$\{T_i\}_1^n$，另记目标关系T_{target}中的目标变量为Y。

定义2 多关系分类的任务即为寻找一个目标分类函数$F(x)$，该函数将每一个目标元组x映射到目标变量$Y=F(x, T_{1,2,\cdots,n}, T_{\text{target}})$，$x\in T_{\text{target}}$。

定义3 两分类数据中AUC[2]为：

$$AUC=\frac{s_0-n_0(n_0+1)/2}{n_0n_1}$$

其中，n_0、n_1分别是样本数据集中正负元组的个数；$s_0=\sum r_i$，r_i表示第i个正例在排序表中的序号。

定义4 T_{target}的非平衡度[3]使用T_{target}中数量最少的类元组与数量最多的类元组数目的比值表示。记非平衡度为B，则$B=\frac{\min\{n_i \mid i=1,2,\cdots,K\}}{\max\{n_i \mid i=1,2,\cdots,K\}}$。其中，$n_i$表示$T_{\text{target}}$中类$y_i$所含的元组数目，$T_{\text{target}}$的非平衡度记为：

$$B=\min\{\frac{\{n_i \mid i=1,2,\cdots,K-1\}}{\sum_{j=i+1,i+2,\cdots,K} n_j}\}$$

显然，$0<B<1$。

基金项目：江苏省高技术研究基金资助项目(BG2007028)；江苏省高校自然科学基金资助项目(09KJB52003)
作者简介：杨鹤标(1960－)，男，教授，主研方向：软件工程，数据挖掘，系统集成；王 健，硕士研究生
收稿日期：2010-04-10 **E-mail：**yhbjj@ujs.edu.cn

3 MMCID 模型与相关算法

MMCID 模型构建过程如图 1 所示，共由 3 个阶段组成：预处理，训练及分类检验。

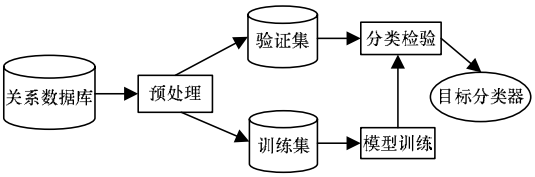

图 1　MMCID 训练及分类过程

3.1　预处理

该阶段为后续训练及验证提供运行时参数，完成下述处理步骤后，可划分训练与验证数据集。

Step1 T_{target} 目标类排序。计算 T_{target} 中每个类 y_i 的出现频率，记为 $freq(y_i)$，由 $freq(y_i)$对类进行排序，可得有序的类列表$<y_1,y_2,\cdots,y_K>$。其中，y_1 是最不频繁的类，y_K 是最频繁的类。

Step2 构造转换矩阵。有效纠错输出编码(Error Correcting Output Coding, ECOC)[4]必须满足 2 个条件：(1)编码矩阵的行之间不相关。(2)编码矩阵的列之间不相关且不互补。因此，对 K 类分类问题，编码长度 L 必须满足 $lbK<L<2^{K-1}-1$。本模型中定义 $L=K-1$，显然当 $K>2$ 时，$lbK<K-1<2^{K-1}-1$，因此，ECOC 编码是适用的。对每个类 y_i 进行长度为 L 的编码记为 $g_i(y_i)$。对$<y_1,y_2,\cdots,y_K>$进行 ECOC 编码，产生行数为 K、列数为 K-1 的转换矩阵 $\boldsymbol{M}$，每个类 y_i 对应长度为 K-1 的二进制编码，则：$\boldsymbol{M}=[g_1(y_1),g_2(y_2),\cdots, g_k(y_k)]^T$。

Step3 建立虚拟连接。在背景关系中建立相对应的 IDset 和 class label，包含 2 种情况：(1)从目标关系 T_{target} 到非目标关系 T_i。(2)从非目标关系 T_i 到非目标关系 T_j。描述如下：

/*T_i 和 T_{target} 可以按属性 T_i.A 和 T_{target}.A 直接连接*/

If　$T_i.A \infty T_{target}.A$

{

/* T_i 中的元组 t 的 IDset(t)即为与元组 t 相连接的目标元组的 ID 的并集*/

$IDset(t)=\cup_{t\in Ti,t.A=Ttarget.A} T_{target.key}$;

/*对于 T_i 中的 class label，若 ID_v 表示 T_{target} 中元组 v 的编号，v 的所属类别为 y_v */

Class label=$\{\sum_{t\in Ti,t.A=Ttarget.A}(y_v=y_1),\cdots,\sum_{t\in Ti,t.A=Ttarget.A}(y_v=y_k)\}$;

}

/* T_i 与 T_j 可以按属性 T_i.A 和 T_j.A 连接*/

If　$T_i.A \infty T_j.A$

{

/* T_i 中的每个元组 t 与 T_{target} 的 ID 集 IDset(t)相关联，则对于 T_j 中的每个元组 u，T_i 中的元组 t 的 IDset 中的所有 ID 都传播到 T_j 中对属性 A 可与 t 连接的每个元组 u*/

$IDset(u)=\cup_{t\in Ti,t.A=u.A}$ IDset(t);

class label=$\{\sum_{t\in Ti,t.A=u.A}(y_v=y_1),\cdots,\sum_{t\in Ti,t.A=u.A}(y_v=y_k)\}$

}

Step4 背景关系聚集。对任一背景关系 T_i 中的属性 A_i(除主键 T_{ikey} 外)，有：(1)A_i 是一个标称型属性，则产生一个新的属性 $\Psi(A_i)$，$\Psi(A_i)$ 的值为运用 COUNT 函数的聚集值。(2) A_i 是一个数值型属性，则将产生 6 个新的属性 $\Psi_{1,2,\cdots,6}(A_i)$，对应属性值依次为运用 SUM、AVG、MIN、MAX、STDDEV 和 COUNT 函数的聚集值。

3.2　子分类器的构造及评估

在训练阶段构造子分类器并对其进行评估验证。依据一对多划分[5]方法，对$<y_1,y_2,\cdots,y_K>$依次构造子分类器，由 CrossMine 共构造 K-1 个子分类器。构造过程为：第 1 个子分类器记为 $F_1(y_1,y_2+)$，该分类器以 y_1 类的元组为正例，y_1 后续所有类的元组为负例进行训练。更一般的，第 i 个分类器记为 $F_i(y_i,y_{i+1}+)$，$1<i<k$，它以 y_i 类的元组为正例，后续所有类的元组为负例进行训练。假设一个数据集含 5 个目标类，其有序的类列表为$<C_2,C_1,C_3,C_4,C_5>$，则第 1 个分类器将 C_2 和 C_1、C_3、C_4、C_5 进行了划分，第 2 个分类器划分了 C_1 和 C_3、C_4、C_5 的边界，其他以此类推。

为保证最终目标分类器的性能，采用 AUC 对子模型进行评估验证。由定义 3 可得各子分类器的 AUC 值。对 AUC 值小于等于 0.5 的子分类器，标记该子分类器状态 $F_i.S$=0，即为不可用状态。这是因为在 ROC 曲线图中，随机猜测算法可在(0,0)和(1,1)产生直线，此类模型的 AUC 值可达 0.5，因此要求子分类器的分类能力至少应好于随机猜测。

3.3　模型分类检验

在检验阶段构造最终目标分类器。经评估验证后的子分类器为 $F_i(1<i<K-1)$，则目标分类器可记为 $F=F_1\cup F_2\cup\cdots\cup F_{K-1}$。如果 $F_i.S!=0$，则 $F_i(x_i)=y_i$ 时，$F_i=1$；$F_i(x_i)\neq y_i$ 时，$F_i=0$。其中，∪代表各子分类器采用的组合策略。一种常用的策略是依据各子分类器预测 x 所属类别的概率进行选择，然而该方法缺少现实意义。这里采用的策略为：对每个子分类器的分类结果作连接产生长度为 K-1 的码字 W，记 $W=\sum_{i=1}^{k-1}F_i(X)$。即判断 x 是否属于类 y_i，若是则为 1，否则为 0。这样由 K-1 个子分类器的输出结果连接形成 K-1 位的二进制码字 W。将 W 与类 y_i 对应的编码 $g_i(y_i)$逐一比较海明距离，定义任一次比较的结果为 $d_i=h(W, g_i(y_i))$，经比较后距离最小的类确定为最终目标类，即 $F(X)=\{y_i|\min\{h(W, g_i(y_i))\}\}$。

图 2 显示了模型分类检验的过程。

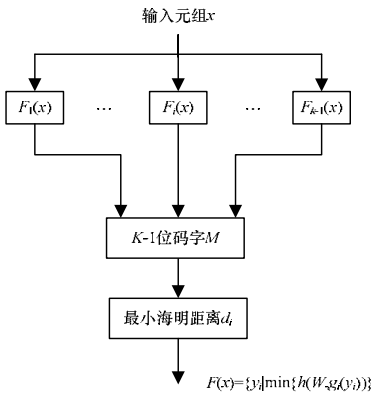

图 2　模型分类预测示意图

在实验部分采用保持和交叉确认的方法进行了分类验证。模型构建算法描述如下：

算法 1 MMCID 算法

输入 数据库 $DB=\{T_{target},T_1,T_2,\cdots,T_n\}$，目标元组$\{<x_i, y_i>\}\in T_{target}$，$x_i\in X$, $y_i\in Y$(X 表示目标元组集合，Y 表示目标类集合)

输出 目标分类模型 F

Sort(Y)→$<y_1,y_2,\cdots,y_K>$;

For $y_i\in<y_1,y_2,\cdots,y_K>$ do $g_i(y_i)$=ECOC(y_i);

End

$M=[g_1(y_1),g_2(y_2),\cdots,g_k(y_k)]^T$;

Propagate(T_{target}, $\{T_i\}_1^n$);

Aggregate($\{T_i\}_1^n$);

Train(T_{target}, $\{T_i\}_1^n$)→$\{F_i\}_1^{k-1}$

If AUC($\{F_i\}_1^{k-1}$)<0.5 F_i.S=0;

$W=\sum_{i=1}^{k-1}F_i(X)$;

For $g_i(y_i)\in M$ do $d_i= h(W, g_i(y_i))$; End

F(X): Min$\{d_i|i=1,2,\cdots,k-1\}\rightarrow y_i$;

4 实验结果及分析

为了检验 MMCID 的分类效率与准确率，在合成数据库和真实数据库中均进行了实验。通过对比分类性能较高的 CrossMine，突出了 MMCID 在非平衡数据下进行多分类任务的特点。

4.1 实验数据

合成数据库由文献[1]中的数据生成器产生。表 1 给出了数据生成器的关键参数。改变$|B|$、C_{min}、C_{max} 3 个参数值，生成多分类的非平衡合成数据集 *RxTyFz*。

表 1 生成器关键参数

名称	描述	默认值
$\|R\|$	关系的个数	x
T	关系中的元组数	y
F	关系中的外键个数	z
C_{min}	目标关系中的最小类数	3
C_{max}	目标关系中的最大类数	10
$\|B\|$	目标关系中的非平衡度	0.12

真实数据库采用某网络学习平台系统的数据集。对应的关系模式如图 3 所示。其中，Student 表为目标关系；Slevel 为目标变量，代表对学生学习能力的划分；其他表均为背景关系。该非平衡数据集包含了 7 个关系表，共计 6 878 条记录，其中目标关系含 500 条记录，目标类含 3 种类别。

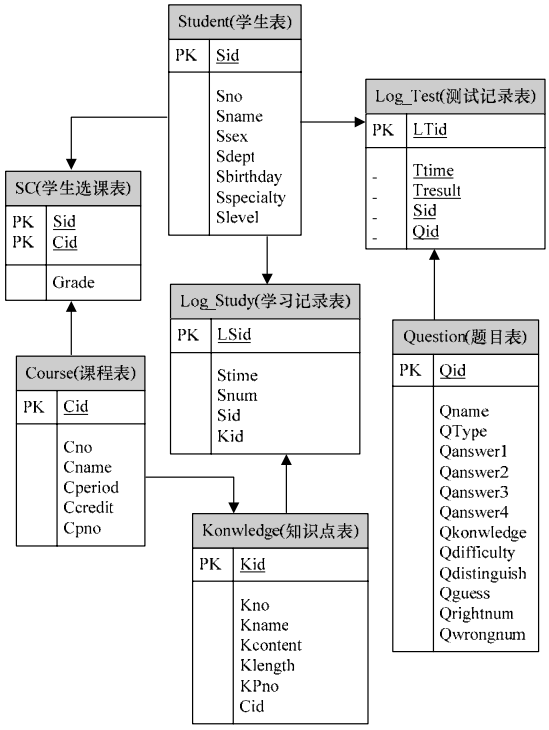

图 3 学习平台系统关系模式

4.2 实验结果

在生成的数据集中前 90%的数据用作训练集，剩下的 10%的数据用作测试集。表 2 列出了模型在生成的 6 个数据库上的测试结果。

表 2 AUC 和运行时间数据

数据集	非平衡度	类数	CrossMine		CrossMine-IM	
			AUC	时间/min	*AUC*	时间/min
R10T1000	0.128	4	0.485	13.08	0.732	18.26
R10T3000	0.096	5	0.422	16.04	0.722	19.38
R10T6000	0.210	6	0.522	15.32	0.842	20.35
R10T10000	0.288	8	0.536	18.38	0.878	22.66
R15T1000	0.144	9	0.496	17.24	0.792	22.86
R20T1000	0.088	10	0.366	20.46	0.703	24.26
Average	0.159	7	0.471	16.75	0.778	21.30

对真实数据库采用了 10 次交叉验证法。图 4 和图 5 显示了随非平衡度的改变，*AUC* 和运行时间对比的情况。

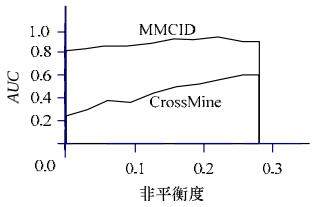

图 4 AUC 对比

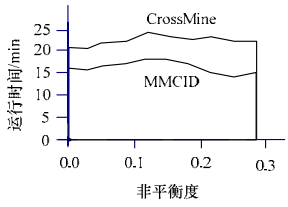

图 5 运行时间对比

4.3 结果分析

研究表明 *AUC* 能定量评估分类器性能，*AUC* 值越大则分类器性能越好。从 2 组实验结果看，在非平衡度越来越高、目标类数量增多的情况下，CrossMine 对应的 *AUC* 值很低，无法达到期望要求，而本模型有较好的分类性能。在运行时间上，由于 CrossMine 无需构造子分类器和分类组合比较，因此运行速度相对较快。

5 结束语

本文将多分类非平衡数据的分类方法推广到多关系分类领域，拓展了多关系分类方法的应用范围。下一步的研究工作可从对子分类器规则集及其间组合策略的优化，在更大规模的数据集上进行验证，提高模型的可伸缩性方面进行展开。

参考文献

[1] Yin Xiaoxin, Han Jiawei, Yang Jiong, et al. CrossMine: Efficient Classification Across Multiple Database Relations[C]//Proc. of the 20th International Conference on Data Engineering. Boston, USA: [s. n.], 2004: 172-795.

[2] Hand D J, Till R J. A Simple Generalization of the Area Under the ROC Curve for Multiple Class Classification Problems[J]. Machine Learning, 200l, 45(2): 171-186.

[3] Murphey Y L, Wang Haoxing, Ou Guobin, et al. OAHO: An Effective Algorithm for Multi-class Learning from Imbalanced Data[C]//Proc. of IEEE International Joint Conference on Neural Networks. Orlando, USA: [s. n.], 2007.

[4] Dietterich T G, Bakiri G. Solving Multiclass Learning Problems via Error-correcting Output Codes[J]. Journal of Artificial Intelligence Research, 1994, 2(1): 263-286.

[5] Rifkin R, Klautau A. In Defense of One-vs-All Classification[J]. The Journal of Machine Learning Research, 2004, 5(12): 101-141.

编辑 顾逸斐