不平衡数据分类方法及其在入侵检测中的应用研究

江 颉[1] 王卓芳[1] GONG Rong-sheng[2] 陈铁明[1]
(浙江工业大学计算机科学与技术学院 杭州 310023)[1]
(美国辛辛那提大学智能系统实验室 辛辛那提 45221)[2]

摘 要 直接将传统的分类方法应用于不平衡数据集时,往往导致少数类的分类精度低下。提出一种基于K-S统计的不平衡数据分类方法,以有效提高少数类的识别率。利用K-S统计评估分类与特征之间的关系,去除冗余特征,并且构建K-S决策树获得数据分片,调整数据的不平衡度;最后对分片数据双向抽样调整,进行分类学习。该方法使用的K-S统计假设条件极易满足,其效率高且适用性强。通过KDD99入侵检测数据的分析对比表明,对于不平衡的数据集,该方法对多数类及少数类都具有较高的分类精度。

关键词 不平衡数据,K-S统计,逻辑回归,入侵检测
中图法分类号 TP181 **文献标识码** A

Imbalanced Data Classification Method and its Application Research for Intrusion Detection

JIANG Jie[1] WANG Zhuo-fang[1] GONG Rong-sheng[2] CHEN Tie-ming[1]
(College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310023, China)[1]
(Intelligent System Laboratory, University of Cincinnati, Ohio 45221, USA)[2]

Abstract The traditional classification algorithms always have low classification accuracy rate especially for the minority class when they are directly employed on classifying imbalanced datasets. A K-S statistic based new classification method for imbalanced data was proposed to enhance the performance of minority class recognition. At first, the K-S statistic was employed as a correlation measure to remove redundant variables. Then a K-S based decision tree was built to segment the training data into several subsets. Finally, two-way resampling methods, forward and backward, were used to rebuild the segmentation datasets as to implement more reasonable classification learning. The proposed K-S based method, with a realistic assumption, is very high efficient and widely applicable. The KDD99 intrusion detection experimental analysis proves that the method has high classification accuracy rate of both minority and majority class for imbalanced datasets.

Keywords Imbalanced data, K-S statistic, Logistic regression, Intrusion detection

1 引言

分类问题是知识发现和数据挖掘领域的一个研究热点。近年来,分类问题中数据的不平衡性越来越受关注。不平衡性指数据集中一类样本的数目远远超过另一类,少数类与多数类的比例达到较低水平。在数据分类算法研究中,往往假定数据对象分布较平衡,但现实世界中观测的数据往往存在不平衡的特点。例如,加拿大银行的贷款产品宣传中,产品推广对象达到90000个客户,但只有1.2%的客户对推广有所反馈,98.8%的客户则没有回应[1]。此外,在医疗诊断、文本分类、入侵检测、欺诈检测等应用领域也都经常出现这类不平衡的数据集。

事实上,数据的不平衡性对分类方法的准确性提出了较大的挑战。一般而言,在极度不平衡的情况下,分类会偏向于多数类的特性,使得多数类的分类正确率较高。相对而言,其较少关注于少数类,少数类的分类结果更容易出现错误。然而,在实际问题中,少数类分类错误比多数类分类错误的后果严重很多。如前面贷款产品推广的例子中,有回应的1.2%客户属于较少的一类,但是其更具有价值,揭示了潜在客户的一些特性。

目前大多数经典的分类算法,如C4.5算法、贝叶斯算法、神经网络等,一般基于如下假设[2]:1)数据集较为平衡;2)少数类与多数类的错分代价一致;3)以整体数据集的分类精度作为评价准则。如果直接将这些方法应用于不平衡数据集,少数类的分类精度将无法保证。为此,针对不平衡数据集的分类问题,本文将提出一种基于K-S统计的新型分类方法,其通过对数据进行分片处理,来调整分片中数据的不平衡性,提高少数类分类精度。

到稿日期:2012-06-13 返修日期:2012-10-18 本文受国家自然科学基金(61103044),浙江省自然科学基金(Y1110567),浙江省科技厅计划项目(2010C31126,2011C21046)资助。

江 颉(1972-),女,博士,副教授,主要研究方向为信息安全、电子服务,E-mail:jj@zjut.edu.cn;王卓芳(1988-),女,硕士生,主要研究方向为数据挖掘、信息安全。

2 研究现状

针对不平衡数据集的分类问题，已经提出了许多解决方法，大致可以分为两类[3]：基于数据层面和基于算法层面。基于数据层面的方法不修改现有算法，通过各种抽样方法调整数据分布，再进行分类学习；基于算法层面的方法则是针对数据集的不平衡性，设计新的算法，或者改良现有算法。

2.1 基于数据的方法研究

通过重抽样、过抽样和欠抽样等抽样方法调整数据集，提高数据集的平衡度。抽样独立于分类方法，因此适用性广。

通过仿真实验，Japkowiz[4]发现随机欠抽样和随机重抽样两种策略都有较好效果，而一些更细致的抽样技术不会产生额外的效果；Chawla[5]研究发现对少数类的过多复制可能会引发过度拟合。因此，他们提出了一种新的重抽样技术SMOTE，即生成一些相邻数据间的值，而不是直接复制已有数据。Drummond 和 Holte[6]使用代价曲线分析抽样效果，结果表明，欠抽样优于重抽样，同时欠抽样对于数据集的变化具有一定的敏感性，且类分布的变化对重抽样几乎不产生影响。另外，还有研究如何通过随机欠抽样来有效减少多数类的方法，例如 Kubat 和 Matwin[7]将多数类分为噪音点、边界点、冗余点、有效点 4 类，通过有效的区分只保留有效点，以此提高分类精度。

总体上，数据抽样在很多领域应用都获得了较好的效果，但是过抽样和欠抽样的方法普遍存在一些缺点：过抽样虚构额外的信息，不仅增加了计算消耗，还可能产生噪音数据或者过拟合现象；欠抽样则可能会丢失一些多数类样本中的重要信息，在增加分类正确的少数类样本数量的同时，对多数类样本的分类会产生负效应。

2.2 基于算法的方法研究

Holte[8]最早提出了决策树中“small disjuncts”的概念，指出由极其少量的样本所产生的规则是分类器误差的主要来源。随后，Weiss[9]指出决策树中将 small disjuncts 与噪音点有效区分存在一定难度；文献[10]通过改进叶节点的评估算法和修剪策略来解决 small disjuncts；文献[11]通过设置错误代价函数，修正对多数类的偏向，降低整体的错误率；文献[12]则采用支持向量机分类器 SVM，调整分类面向多数类方向移动，从而使得少数类样本的分类精度得到一定的提高。

一般来说，算法改进可获得较好效果，但由于具有模型针对性，通常无法直接将其应用于其它不平衡数据。

综上所述，抽样和改进分类方法均能够提高少数类的分类正确率，但各有其缺陷。此外，传统的特征选择方法如信息增益、χ^2 统计方法、相关系数方法等，通过一个评价函数评估特征，也并不适用于不平衡数据处理[13]。目前，各种分类算法通常依据整体数据集，将问题划分为若干子问题。这样的划分极易偏向多数类，造成少数类样本的数据碎片。

鉴此，本文提出一种基于 K-S 统计量的不平衡数据集分类方法，其在保证少数类和多数类的分类精度的同时，具有适用性。

3 本文的方法

3.1 K-S 统计方法

统计学中，K-S(Kolmogorov-Smirnov)检验法用来检验单一样本是否服从某一预先假设的特定分布(单样本 K-S 检验)，通过将样本数据的累积分布函数与理论分布函数相比较来建立统计量。

双样本 K-S 检验被广泛用来检验两个样本概率分布是否相同。若初始假设两个样本的累积分布函数是一致的，则 $\sqrt{\frac{N_1N_2}{N_1+N_2}}D>K_\alpha$ 时，假设被拒绝，D 为两个样本的经验累积分布曲线之间的最大距离，即 K-S 统计值，N_1 和 N_2 分别为两个样本的大小，K_α 由 $P(K<K_\alpha)=1-\alpha$ 确定，K 是服从 Kolmogorove 分布的一个随机值。

当 K-S 统计被用来衡量输入变量(即某个特征属性)与输出分类之间的关系时，可以如下定义 K-S 值：

$$KS=\max_{x\in S}|F^+(x)-F^-(x)| \tag{1}$$

F^+ 和 F^- 分别是某一特征正类样本和负类样本的经验累积分布函数，S 为此输入变量所有可能的取值。K-S 统计值越大，正类和负类的累积分布函数越不一致，即输入变量能够有效区分正类和负类样本，输入变量与输出分类之间的相关性越强。

K-S 统计具有以下几个优点。1)K-S 统计只需计算相关频数，所以数据的不平衡性对 K-S 值不会产生影响；2)不同离散化方法会产生不同的离散结果，从而影响分类效果，而 K-S 统计无需离散化，避免了此类误差影响；3)对于大多数的关系型分类方法，K-S 统计无需任何统计假设，适用性广。

3.2 逻辑回归分类

逻辑回归是一种用于描述分类型回应变量与一组预测变量之间关系的方法[14]。在线性回归中，给定初始条件 $X=x$，回应变量 Y 是一个随机变量，定义为：

$$Y=\beta_0+\beta_1x+\varepsilon \tag{2}$$

其条件均值即线性回归线为 $E(Y|x)=\beta_0+\beta_1x$。逻辑变换后，逻辑回归的条件均值表现形式如下：

$$\pi(x)=\frac{e^{\beta_0+\beta_1x}}{1+e^{\beta_0+\beta_1x}},0\leqslant\pi(x)\leqslant1 \tag{3}$$

$\pi(x)$亦即在 $X=x$ 条件下事件 Y 发生的概率 $E(Y|x)$。逻辑变换将评估概率限制在 0～1 之间，用来构造二分变量模型，具有很好的灵活性和可解释性。

估计逻辑回归系数时，通常采用最大似然估计法。正向效应的概率为 $P(Y_i=1|X=x_i)=\pi(x)$，负向效应为 $P(Y_i=0|X=x_i)=1-\pi(x)$。因此，$Y_i=0$ 或 1，对第 i 个观察值概率的影响可以表示为 $[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}$。样本独立情形下，似然函数可以表示为：

$$l(\beta|x)=\prod_{i=1}^{n}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i} \tag{4}$$

通过迭代计算，使似然函数最大化，可得到回归参数。

3.3 不平衡数据分类

3.3.1 方法总体框架

本文提出一种系统的分类方法，总体流程如图 1 所示。

首先使用 K-S 统计对数据进行预处理，完成特征选择，去除不相关特征，减轻分类的计算压力；基于 K-S 统计结果，构建 K-S 决策树，最终选定特征和数据分片，其中决策树终止条件不唯一，K-S 值或 K-S 值的变化率都可作为阈值，或依据问题的具体情况综合考虑；接下来对每个分片进行双向抽

样调整：首先对多数类欠抽样，样本最优大小由后向搜索算法决定，接着对少数类抽样，抽样程度由前向搜索算法确定；最后在每个分片上进行逻辑回归分类。

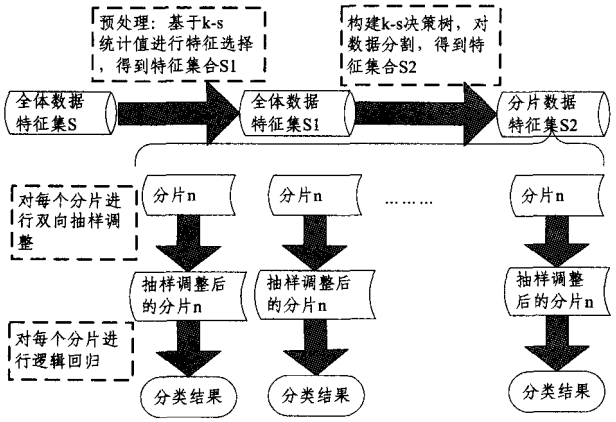

图 1　方法总体框架

3.3.2　数据分片

对数据进行分片，即将一个比较大的数据集划分为几个较小的数据集，使得相似的数据集中在某个较小的数据集中。数据分片不仅降低了整体的计算复杂度，也使得分片中的数据特性更集中。在不平衡数据分类问题中，少数类中存在小集团的现象，一些数据集中于某一个较小的分布，存在于某一个分片中，对其影响最大的特征也各不相同。

下面介绍基于 K-S 决策树的数据分片步骤。将第一次特征选择选取的特征作为构建 K-S 决策树的备选特征。首先在备选特征中进行 K-S 统计，最大 K-S 值特征具有最大分类能力，将其分界点作为数据集的划分点。然后对所有子集进行 K-S 统计，选取生成最大 K-S 统计值的点作为划分点，划分其所在子集。如此递归，直到最大 K-S 值低于设定的阈值，或分片数量超过一定值。算法具体描述如下。

Step1　数据集 $D(F,C)=D(F_1,F_2,\cdots,F_n,C)$ 有 n 个特征属性 F_1，$F_2,\cdots,F_n$，C 为类别属性，γ 为决策树终止的判定条件。

Step2　While (Notγ) do

Step2.1　对于数据集 D 的每个子集 D_i，计算其中每个特征 F_j 的 K-S 值 KS_{ij}，划分点记为 $Split_{ij}$；

$$[KS_{ij},Split_{ij}]=KSSplit(F_{ij},C_i) \tag{5}$$

Step2.2　选取最大的 K-S 值记为 $KS_{i'j'}$，则最终即对子集 $D_{i'}$ 进行划分，划分特征为 j'，划分点为 $Split_{i'j'}$；

Step2.3　将 Step2.2 划分得到的子集视为独立的数据集，重复 Step2.1。

一般而言，每个类的概率分布函数都是单调的，则不论其概率分布函数是何类型，两个概率分布曲线的交叉点即为两个类累积密度函数差异值最大的点，也就是 K-S 值最大的点，因此可选择这个点作为决策树的划分点。由于这个假设十分宽松，且符合实际应用，因此这样的选择方法具有较高的健壮性。

3.3.3　数据重采样

数据分片面临的最大困难是如何确定类分布的最优情形。研究发现，类分布没有普适的最优处理方法，而与实际问题相关[3]。一般重抽样和欠抽样都能够调整数据，使之更平衡。因此，我们采取双向抽样调整来确定最优分布。具体方法如下：首先对多数类进行欠抽样，采取不放回的随机抽样策略，即后向搜索策略，将欠抽样初始比例设为 100%，每次迭代中，按预先设定的步长减少比例，每次迭代结束都对数据进行模型检测。当分类效果明显降低时停止迭代，产生的最优效果的样本比例就是欠抽样的最优比例。重抽样的方法与欠抽样相似，区别在于采取放回的随机抽样，使用前向搜索策略，即初始比例设为 100%，在每次迭代中，按一定步长增加比例。当模型检测效果降低时，停止迭代，此时产生的最优效果的比例就是重抽样的最优比例。

3.3.4　算法应用实例分析

下面以 KDD99 数据集为例，给出不平衡度为 3%的二分问题分类过程。KDD99 提供了网络异常检测的标准数据测试集，10%数据集的样本总数为 50 万条，每个数据包含一个 TCP/IP 网络连接提取出的 41 维特征和 1 维类标记。实验数据集中异常类标记的是各种子类异常名称，因此在对数据集进行预处理过程中需要将子类异常统一标记为异常。根据实验需要，从中随机抽取正类样本 1500 条、负类 48500 条构造成不平衡度为 3%的数据集。

首先，对数据进行预处理，利用 K-S 统计分析每个属性与分类之间的关系，挑选相关特征。对于类别型变量，可以针对其每个值构建一个二值变量，与连续性变量一样计算 K-S 值。取一系列二值变量中最高的 K-S 值作为这个变量的 K-S 统计值，如特征 x2 为类别型变量，含有 3 个值，分别为 tcp，udp，icmp。这 3 个值所具有的 K-S 统计量分别为 0.5036，0.1929，0.6965，则特征 x2 的 K-S 统计值为 0.696。K-S 值最高的 15 个特征如表 1 所列。

表 1　不平衡度 3%数据集各项特征的 K-S 统计值排名

特征	描述	类型	K-S 统计值	排名
x23	duration	continuous	0.9671	1
x6	protocol_type	symbolic	0.8431	2
x12	service	symbolic	0.7155	3
x24	flag	symbolic	0.7078	4
x3	src_bytes	continuous	0.7033	5
x2	dst_bytes	continuous	0.6954	6
x36	land	symbolic	0.6626	7
x32	wrong_fragment	continuous	0.6030	8
x5	urgent	continuous	0.5808	9
x37	hot	continuous	0.5294	10
x31	num_failed_logins	continuous	0.3347	11
x29	logged_in	symbolic	0.2670	12
x30	num_compromised	continuous	0.2410	13
x26	root_shell	continuous	0.2187	14
x39	su_attempted	continuous	0.2187	15

取 0.5 作为 K-S 统计值的阈值，去除统计值小于 0.5 的特征，只保留前 10 个特征作为第一次特征选择的结果：x23，x6，x12，x24，x3，x2，x36，x32，x5，x37。接着由筛选出来的特征构建 K-S 决策树，对数据进行分割，决策树中的关键特征即为第二次特征选择的结果。决策树阈值条件取分片次数为 10。第一次划分中，特征 x23 值为 62 时，得到最大 K-S 统计值，即取 x23=62 作为划分点，将数据分为 2 片。

特征 x23 为连接时间，因此连接时长是否大于 62 是判定连接是否正常的一个重要依据。当分片次数达到 7 时，分片已经完成，划分点不再变化。最终的决策树如图 2 所示。

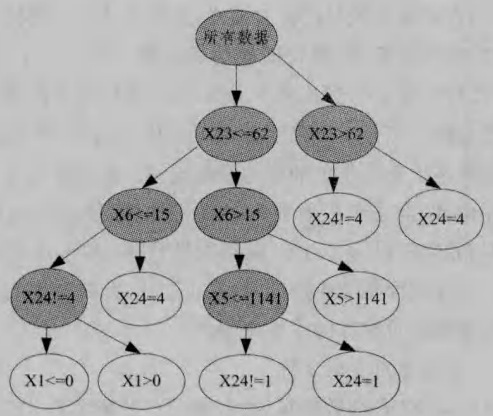

图 2　不平衡度 3%的 K-S 决策树

数据分片过大会影响数据的聚集，造成数据碎片、过拟合等现象。因此本实验中调整决策树为 3 个分片，如图 3 所示。对每个分片双向抽样调整后，采用逻辑回归进行分类学习。

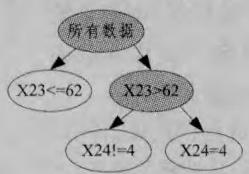

图 3　3 个分片的 K-S 决策树

4　入侵检测实验

4.1　评价指标

本文实验采用 *F-value* 来评价不平衡数据集的分类性能。对于 2 类分类问题，*F-value*[15] 定义如下：

$$F\text{-}value=\frac{(1+\beta^2)\times Recall\times Precision}{\beta^2\times Recall\times Precision} \tag{6}$$

式中，*Recall* 为查全率，*Precision* 为查准率，β 为 *Precision* 与 *Recall* 的权重比例，可取值为 1。

$$Recall=\frac{TP}{TP+FN},Precision=\frac{TP}{TP+FP} \tag{7}$$

式中，*TP*，*FN*，*FP* 是由混淆矩阵所定义的值。

采用 *F-value* 而不是常用的分类精度，是因为对于不平衡数据集分类精度无法正确评判少数类的分类准确率。如果将所有样本都分为多数类，那么在多数类达到 99%时，整个分类结果的分类精度也达到了 99%。但是这样的结果是不合理的，少数类的分类正确性无法保证，其分类意义无法达到。而由 *F-value* 的定义可得，只有在查全率和查准率都较大时，*F-value* 值才会较大。因此 *F-value* 对于不平衡数据集更为适用。

4.2　实验分析

本实验数据集选用 KDD99(10%)数据集，共 50 万条数据，其中负类数据为 396742 条，正类为 97278 条。为了观察本文提出的方法对不同不平衡程度的数据集的效果，分别从正类和负类中随机抽取数据构建多个测试数据集，使得每个数据集的样本数量为 50000 条，且正类数与负类数的比例分别为 1%，3%，5%，8%，10%，12%，15%；同时对分类标记做预处理，标记为正常与异常 2 类。

分片数量确定数据聚集程度，最终影响分类精度。下面首先以不平衡度 3%为例，来分析分片数量对分类的影响。表 2 为不平衡度 3%数据集不同分片数量下 *F-value* 的值。当分片数为 4 片时，*F-value* 值急剧下降。观察各个分片中的数据，发现第 4 个分片为空分片，第 4 个划分点为所在分片的边界点。5 个分片下情况相同。对于本数据集，3 个分片为最优分片数，此时 *F-value* 值最大。

表 2　不同分片数下 *F-value* 值对比

分片数	2	3	4	5
F-value 值(%)	39.71	38.97	8.69	25.74

数据分割调整了每个分片上的样本不平衡度，观察表 3 可知，分片 a 中的不平衡度上升至 55.92%，分片 b 中所有的数据都是负类，分片 c 中所有的数据都是正类。因此逻辑回归中，在分片 b 和 c 中的数据，正类的概率将分别为 0 和 1，这两个分片也无需进行抽样调整。对于分片 a，在逻辑回归前进行数据重采样，对数据进行进一步的不平衡度调整，即 150%的过抽样和 75%的欠抽样。采用十折交叉法对数据进行逻辑回归，最终不平衡度 3%的 *F-value* 值为 43.62%。

表 3　不平衡度 3%的分片结果

分片	描述	样本数	样本比例(%)	正类样本数	正类样本比例(%)
A	X23<=62	2659	5.32	1487	55.92
B	X23>62,X24!=4	47328	94.66	0	0
C	X23>62,X24=4	13	0.02	13	100
Total		50000	100	1416	3

下面采用本文的方法对各个数据集进行分类，与直接使用逻辑回归分类作对比。在十折交叉验证实验环境下的分类结果如表 4 所列。

表 4　不同平衡度下的分类性能比较

不平衡程度	1%	3%	5%	8%	10%	12%	15%
逻辑回归(%)	13.76	12.06	11.29	10.76	10.26	10.51	11.10
分片的逻辑回归(%)	73.78	43.62	32.31	30.97	27.52	23.62	21.96

实验效果如图 4 所示，可以看出，逻辑回归在数据不平衡的情况下，其分类效果不是十分理想，保持在 10%左右这个较低的水平。对于不同不平衡程度数据集，本文提出的方法其分类精度均高于直接采用逻辑回归的方法。数据不平衡程度越高，基于分片的逻辑回归提升效果越明显，如当不平衡程度达到 1%时，*F-value* 值可以从 13.76%提升至 73.78%。因此本文提出的方法能够有效提高各种不平衡程度下数据的分类效果。随着数据不平衡程度从 1%降低到 15%，基于分片的逻辑回归分类精度不断降低，当不平衡程度为 15%时，*F-value* 值只能从 11.10%提升至 21.96%。因此本文提出的方法适用于不平衡程度较高的情况；当不平衡程度较低时，分片依旧能够提高分类效果，只是提升效果不明显。

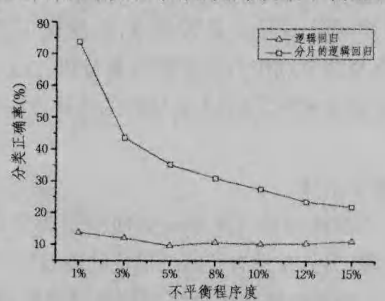

图 4　逻辑回归与分片的逻辑回归分类性能对比图

结束语 本文提出了一种基于 K-S 统计的不平衡数据分类方法，该方法通过对数据分片调整数据不平衡度，然后进行分类学习。分片结果表明，多数类与少数类达到了很好的聚集。将多数类或少数类集中于某一分片，或者在分片中使正类与负类的分布差异性最大，可便于分类器区别。同时对于不同程度的不平衡样本，该方法的分类精度有一定程度的提高。可见，本文提出的方法是有效、可行的。本文对于极端不平衡的数据集具有很好的效果，但是如何进一步提高不平衡程度一般的数据集，将是今后需要进一步研究的目标。分片数量对分类精度有所影响，如何自适应地确定决策树的大小也是今后的研究任务。

参 考 文 献

[1] Ling C X, Li C. Data mining for direct marketing: Problems and solutions[C]// Proceedings of the 4th international conference on knowledge discovery and data mining. New York, NY, 1998: 73-79

[2] Sun Yan-min, Kamel M S, Wong A K C, et al. Cost-Sensitive Boosting for Classification of Imbalanced Data [J]. Pattern Recognition, 2007, 40(12): 3358-3378

[3] Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets [J]. Computational Intelligence, 2004, 20(1): 18-36

[4] Japkowicz N, Stephen S. The class imbalance problem: A systematic study [J]. Intelligent Data Analysis, 2002, 6(5): 429-450

[5] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling techniques [J]. Journal of Artificial Research, 2002, 16: 321-357

[6] Drummond C, Holte R C. C4. 5, Class imbalance, and cost sensitivity: Why under-sampling beats over-sampling [C]// Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets. 2003

[7] Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection [C]// Proceedings of the 14th International Conference on Machine Learning. 1997: 179-186

[8] Holte R C, Acker L E, Porter B W. Concept learning and the problem of small disjuncts[C]// Proceedings of the 11th joint international conference on artificial intelligence. 1989: 813-818

[9] Weiss G M. Mining with rarity: A unifying framework [J]. ACM SIGKDD Explorations Newsletter-Special Issue on Learning from Imbalanced Datasets, 2004, 6(1): 7-19

[10] Quinlan J R. Improved estimates for the accuracy of small disjuncts [J]. Machine Learning, 1991, 6(1): 93-98

[11] Ling C X, Sheng V, Yang Q. Test strategies for cost-sensitive decision trees [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(8): 1055-1067

[12] Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines [C]// Proceedings of international joint conference on artificial intelligence. 1999: 55-66

[13] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced Data [J]. SIGKDD Explorations, 2004, 6(1): 80-89

[14] Larose D T. 数据挖掘方法与模型[M]. 北京：高等教育出版社，2011：143-146

[15] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A New Over-Sampling Method in imbalanced Data Sets Learning[C]// Proceedings of the International Conference on Intelligent Computing. Hefei, China, 2005: 878-887

(上接第 130 页)

[8] Hong Liang-jie, Davison B D. Empirical study of topic modeling in Twitter[C]// Proceedings of the First Workshop on Social Media Analytics. Washington DC, USA, 2010: 80-88

[9] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]// Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press Arlington, Virginia, United States, 2004: 487-494

[10] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for information discovery[C]// Proceedings of the Tenth ACM SIGKDD International Confe-rence on Knowledge Discovery and Data Mining. Seattle, WA, USA, 2004: 306-315

[11] Ramage D, Dumais S, Liebling D. Characterizing micorblogs with topic models[C]// Proceedings of the 4th International Conference on Weblogs and Social Media. Washington DC, U S A, 2010

[12] Daud A, Li Juan-zi, Zhou Li-zhu, et al. Exploiting temporal authors interests via temporal-author-topic modeling[C]// Proceedings of 5th International Conference on Advanced Data Mining and Applications. Verlag Berlin, Heidelberg, 2009: 435-443

[13] Liu Yan, Niculescu-Mizil A, Gryc W. Topic-link LDA: joint models of topic and author community[C]// Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, QC, Canada, 2009: 665-672

[14] Wang Xue-rui, McCallum A. Topics over time: a non-markov continuous-time model of topical trends[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 424-433

[15] McCallum A, Corrada-Emmanuel A, Wang Xue-rui. Topic and role discovery in social networks[C]// Proceedings of 19th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA: 786-791

[16] Mei Qiao-zhu, Liu Chao, Su Hang, et al. A probabilistic approach to spatiotemporal theme pattern mining on weblogs[C]// Proceedings of the 15th International Conference on Word Wide Web. Edinburgh, Scotland, UK, 2006: 533-542

[17] Su Yi-zhou, Han Jia-wei, Gao Jing, et al. iTopicModel: Information Network-Integrated Topic Modeling[C]// Proceeding of the 9th IEEE International Conference on Data Mining. Miami, USA, 2009: 487-497

[18] Mei Qiao-zhu, Cai Deng, Zhang Duo, et al. Topic modeling with network regularization[C]// Proceeding of the 17th International World Wide Web Conference. Beijing, China, 2008: 101-111