基于欠采样支持向量机不平衡的网页分类系统①

李村合, 唐 磊

(中国石油大学 计算机与通信工程学院, 青岛 266580)

摘 要: 在这个信息爆炸的时代, 如何处理这些海量的数据如何有效的分类已经引起了人们的高度重视, 尤其是在互联无技术迅速发展的阶段, 网页分类这领域已成为热点. 与传统的分类方法相比, 支持向量机具有高维、小样本、适应性强的特点, 能够非常有效率的解决网页分类问题, 但是不平衡数据的分类这一方面, 存在着分类不精确的问题. 所以本文提出了新的解决不平衡数据样本策略, 便是将欠采样策略与传统的支持向量机结合起来, 在减少多数类样本集中噪声数据的基础上增加少数类的样本集数量, 从而使得不平衡样本集趋向于平衡, 最后结合 SMO(Senquential Minimal Optimization)算法改进分类器, 提高了分类的准确性.

关键词: 支持向量机; SMO 算法; 训练集缩减算法; 网页分类; 多类分类

Realization of Web Page Classificationn System Based on Under-Sampling Support Vector Machine

LI Cun-He, TANG Lei

(College of Computer and Communication Engineering, China University of Petroleum, Qingdao 266580, China)

Abstract: In this era of information explosion, how to handle these vast amounts of data and how to classify the data effectively has attracted much attention, especially in the stage of rapid development of Internet technology free, the field of web classification has become a hot spot. Compared with the traditional classification methods, support vector machine has the characters of high-dimensional, small sample size, strong adaptability, and can be very effective to solve the problem of web page classification. But in the field of classification of imbalanced data, there is a problem of inaccurate classification. Therefore, this paper proposes a new strategy to solve the imbalance data samples, that is, combining the under-sampling strategy with the traditional support vector machines to increase the number of samples set in the minority class and to reduce the concentrated noise data in the majority class, so that imbalanced sample set tends to be balanced. Finally SMO algorithm is used to improve the accuracy of classification.

Key words: support vector machine; SMO algorithm; reduction in the training set; classification of web page; multi-class classification

1 引言

随着互联网的发展, 网络的信息量爆炸式的增长, 人们从互联网中获取有用的信息越来越困难, 也给互联网的企业带来的挑战, 由此, 网页分类技术如雨后春笋般发展起来, 在众多的分类方法中, 支持向量机具有较强的学习能力, 经成为网页分类界研究的热点尤其是搜索引擎领域[1]. 在分类的算法里, 支持向量机是非常优秀的算法, 支持向量机是一统计学理论和结构风险最小化原则为基础的学习机器[2], 在分类领域具有非常广泛的应用, 在平衡问题的表现上效果非常好, 可以克服局部最小值问题, 但是在支持向量机在处理不平衡样本集时其分类效果并不理想[3]. 出现上述问题的原因是传统的分类算法都是以提高整体分类准确率作为目标, 然而, 实际问题中存在大量的不平平衡样本集: 某一类的样本数量远少于其他样本数量, 例如信用卡欺诈行为检测, 网络入侵行为检测医

① 收稿时间:2016-07-09;收到修改稿时间:2016-08-08 [doi: 10.15888/j.cnki.csa.005671]

学疾病诊断[4], 产品合格检查. 不平衡样本集的分类问题是, 总体准确率还可以, 然而少数样本的分类准确率很低, 可是在很多的实际问题中, 比如在癌症检测中, 健康细胞现相对于癌细胞是多数类, 对癌细胞的正确分类更重要[5], 在辨别垃圾邮件中, 垃圾邮件是少数类, 然而对垃圾邮件的辨别更重要. 在众多解决支持向量机不平衡分类的研究中, 比较出色的的策略是欠采样算法, 和修改核函数[6]. 本文绍了支持向量机, 并且在传统的支持向量机中结合欠采样策略在一定程度上解决了不平衡网页分类的问题.

2 相关技术

2.1 支持向量机

最早, 是由 Vapnik 提出支持向量机(SVM)理论[7]. 该方法采用结构风险最小化原则代替传统经验啊风险最小化原则, 可以使其在训练样本数量有限的情况下, 很好的兼顾分类识别准确率和分类推广能力, 因此 SVM 被广泛的运用到模式识别和分类问题中[8]. 其原理如下:假设存在一个具有两类样本的训练样本集合如公式(1)所示:

$$D=\{(x_1,y_1),\ldots,(x_l,y_l)\},x\in R^n,y\in\{+1,-1\} \quad (1)$$

当线性可分时则有如下公式(2)所示超平面:

$$\langle w,x\rangle+b=0 \quad (2)$$

其中, ω 代表的是该超平面的法线的方向.

求解最优分类超平面 $\langle w,x\rangle+b=0$, 即在训练集一定的情况下, 确定参数 ω 和 b 的最佳值, 以最小化下面的公式(3):

$$\varphi(\omega)=\frac{1}{2}\|\omega\|^2 \quad (3)$$

同时满足约束条件:

$$y_i[<w,x_i>+b]\geq 1,i=1,\cdots l \quad (4)$$

该问题的求解是一个寻优问题, 在此为凸二次规划问题[9], 通过引进拉格朗日算子 $\alpha_i\geq 0,i=1,\ldots,l$, 得到如下拉格朗日公式:

$$L(\omega,b,\alpha)=\tfrac{1}{2}\|\omega\|^2-\sum_{i=1}^{l}\alpha_i\{y_i[<\omega,x_i>+b]-1\} \quad (5)$$

这里 L 的极值点为鞍点, 可取 L 对 α 的最大值 $\alpha=\alpha^*$, 和对 ω、b 的最小值 $\omega=\omega^*$, $b=b^*$, 将原问题的求解最小化问题转换成求解最大化问题, 满足约束条件:

$$\sum_{i=1}^{l}y_i\alpha_i=0,0\leq\alpha_i\leq C,i=1,\ldots,l \quad (6)$$

解如下问题:

$$\begin{aligned}\max W(\alpha)&=\sum_{i=1}^{l}\alpha_i-\tfrac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{i}\alpha_i\alpha_i y_i y_i K<x_i,x_j>\\&=e^T\alpha-\tfrac{1}{2}\alpha^T Q\alpha\end{aligned} \quad (7)$$

得到最优超平面:

$$\omega^*=\sum_{i=1}^{m}\alpha_i y_i x_i \quad (8)$$

$$b^*=-\tfrac{1}{2}<\omega^*,x_r+x_s> \quad (9)$$

其中, x_r, x_s 为任意支持向量. 从公式(8)中可以发现, 当一个样本的 $\alpha_i=0$ 时, 它对分类没有任何影响, 只有当 $\alpha_i>0$ 才会影响到 ω* , 进而对分类结果产生影响分类, 这种对分类结果有影响的样本称之为支持向量.

相应的分类器为:

$$f(x)=\mathrm{sgn}(<\omega^*,x>+b^*)=\mathrm{sgn}\{\sum_{i=1}^{m}\alpha_i^* y_i K<x,x_i>+b^*\} \quad (10)$$

对于一个未知类别的样本 x, 若 $f(x)=+1$, 则样本 x 属于正类, 即当前类别, 若 $f(x)=-1$, 则样本 x 属于负类, 即非当前类别.

2.2 SMO 算法

SMO 算法的全称是 Senquential Minimal Optimization, 是由 John C. Platttu 提出来的[10]. 支持向量机训练的问题实际上是一个凸优化问题, 即求解一个受约束的二次规划(quadratic programming, QP)问题, 其中比较有名的就是 SMO 算法, SMO 算法的思想是循环迭代: 都是将原有大规模的 QP 问题分解成一系列小的 QP 问题, 按照某种迭代策略, 反复求解小的 QP 问题, 构造出原有的大规模 QP 问题的近似解, 并使该近似解逐渐收敛到最优解[11].

3 基于欠采样不平衡SVM算法实现

针对不平衡数据集上使用传统支持向量机的时候, 少数类的分类精度非常低. 最近有许多解决不平衡 SVM 分类的方法, 这些解决方法主要是从数据和算法方面入手[12]. 从数据这方面入手, 主要有重采样方法, 随机下采样方法, 向下采样方法[13]. 把训练数据集中的的多数类样本减少, 以提高分类性能, 被称为欠采样方法[14]. 为了减少样不平衡样本集中多数类, 本文对多数类样本集进行欠采样处理, 于是提出了基于欠采样不平衡 SVM 算法(Under-sampling-SVM).

下面是算法的定义:

定义 1. (距离)给定两个数据集, x_i, x_j, 在特征空间

上两个样本之间的距离可以表示为:

$$d(x_i,x_j)=\sqrt{K(x_i,x_i)-2K(x_i,x_j)+K(x_j,x_j)} \quad (11)$$

定义 2. 这些发样本的平均特征成为这些样品的中心. 给定训练样本 $\{x_1,x_2,...,x_n\}$, 所以这些样本的中心可以表达为:

$$c=\frac{1}{n}\sum_{i\in X}\varphi(x_i) \quad (12)$$

如果使用不平衡训练数据集, 传统SVM分类分级后, 将接收到的样本的类别是相同的少数类训练数据集的有准确的分类. 每当在不平衡数据样本集上采用传统的SVM进行分类, 得到的超平面会靠近少数类的侧面. 分类后, 少数类的几个样品被分类为多数, 一边, 多数类的样品几乎列为多数的侧面. 因此少数类的分类精度不高. 如示于图 1.

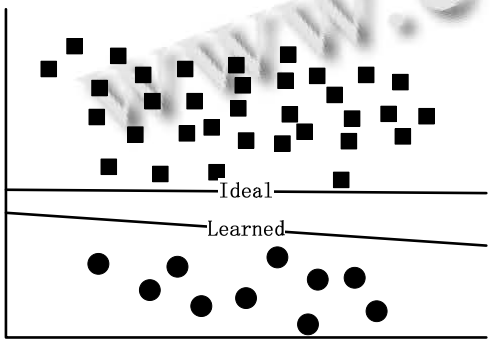

图 1　不平衡训练样本集表现

根据支持向量机的几何特征确定如何减少大多数类的不平衡支持向量机学习的大小应该考虑一个样本的大多数类和少数类的中心之间的距离. 基于定义 1 的结果, 增加不平衡支持向量机学习中的少数类的大小应该考虑不平衡支持向量机分类的结果因此, 当做不平衡的支持向量机学习时, 我们使用下面的筛选机制: 尽我们所能选择的样本, 大多数类封闭的少数类的中心作为新的大多数样本, 增加少数类的样本添加到少数样本的不平衡支持向量机分类.

所以, 基于欠采样不平衡 SVM 算法可以概括如下: (设置原始不平衡样本集作为一个 A ,设置少数类样本 A_+, 设置多数类样本集为 A_-)

本文使用的是核函数为: 高斯核函数(RBF)

第一步: 在多数类样本集 A_- 上, 通过欠采样获得一个新的样本集 A_{new1}. 首先, 如果 A_- 和 A_+ 比例超过了 2, 对多数类样本集 A_- 进行欠采样, 得到一个新的多数类样本集 A_{-new1}, (保持 A_{-new1} 和 A_+ 比例为 2)那么, 新的样本集 $A_{new1}=A_+\cup A_{-new1}$. 那么 A_- 和 A_+ 中心的距离表示为: 根据公式(2), A_+ 的中心表示为:

$$c_+=\frac{1}{n_+}\sum_{i\in A_+}\phi(x_{+i}) \quad (13)$$

其中 x_{+i} 是 A_+ 中的一个样本. 通过公式(1), A_- 中的一个样本 x_{-i} 和 A_+ 的中心 c_+ 距离为:

$$d_{c_+x_{-i}}=\sqrt{K(x_{-i},x_{-i})-2K(x_{-i},c_+)+K(c_+,c_+)} \quad (14)$$

第二步: 对于经过欠采样处理的样本集 A_{new1}, 通过使用 SMO 算法对其进行训练[15],从而的得到一个训练模型 M_1, 即得到一个分类器(分类函数), 如2.1所讲

$$f(x)=\mathrm{sgn}(<\omega^*,x>+b^*)=\mathrm{sgn}\{\sum_{i=1}^{m}\alpha_i^*y_iK<x,x_i>+b^*\} \quad (15)$$

第三步: 通过使用得到的模型(分类器) M_1 对测试集 B 进行分类, 测试样本的分类结果为 B_+ (和 A_+ 标记一样, 为少数类), B_- (和 A_- 的标记一样, 为多数类)

第四步: 获得新的训练样本集 A_{new2}. 首先, 往 A_+ 添加 B_+, 于是我们就可以得到一个新的少数类样本集 A_{+new2}, 那么 $A_{+new2}=A_+\cup B_+$. 接下来, 通过对少数类样本集 A_- 进行向下采样得到一个新的样本集 A_{-new2} (保证 A_{-new2} 和 A_{+new2} 的比率为 2), 那么 $A_{new2}=A_{+new2}\cup A_{-new2}$, 向下采样的方法和第一步相同.

第五步: 通过 SMO 算法对新得到的样本集 A_{new2} 进行训练, 从而得到一个新的训练模型 M_2, 即得到一个新的分类器.

第六步: 利用这新得到的模型 M_2 对测试集进行分类.

4　实验结果

大多数不平衡学习的研究更关注两个类别, 因为多分类问题可以简化为两类. 根据惯例, 少数样本被标记为正类, 而大多数样本被标记为负类. 通常应用两个评价标准, 准确率, 召回率[16], 评估不平衡数据集上的分类. 他们描述如下:

$$\text{准确率}=\frac{TP}{TP+FP},\quad \text{召回率}=\frac{TP}{TP+FN}$$

TP(真正类), TN(真负类), 分别代表正类和负类样本的数量要正确分类的次数. FN(假负类), FP(假正类), 分别代表被错误分类的样本数. 他们描述如下. 表 1 代表着两类矩阵.

表 1 两类矩阵

	预测正	预测负
真正类	TP(真正类)	FN(假负类)
真负类	FP(假正类)	TN(真负类)

为了测试 Under-sampling-SVM 算法的性能, 我们将该算法应用到网页分类. 我们比较 Under-sampling -SVM 算法下采样对不平衡数据集传统的 SVM 算法和随机抽样 SVM. 所有的算法都是在用 libsvm-3.1 工具包进行. 三种算法的简单描述如下:

传统 SVM: 在支持向量机训练之前没有任何操作.

随机抽样 SVM: 在该算法中, 一些训练样本被随机删除之前, 支持向量机训练.

Under-sampling-SVM: 该算法不仅消除了一些样品(非支持向量)的基础上的距离, 也提高了少数类反馈机制.

在这实验中, 我们使用了四种不同的 UCI 数据集[17], 包括 Iris, Abalone, Balance Scale and Yeast 数据集. 在四个 UCI 数据集, 一种是选择作为一个少数类和其余类型分为多数类. 每个数据集被随机分为训练集和测试集. 还我们给他们不同的不平衡比(少数类与多数类的比率), 描述 4 个数据集具体况如表 2 所示.

表 2 数据集详情

Dataset	实例数目	多数类样本集	少数类样本集	测试样本集	不平衡比率
Iris	150	66	33	51	2:1
Abalone	4177	3000	300	877	10:1
Balance Scale	625	400	33	192	12.1:1
Yeast	1484	990	33	461	30:1

UCI 数据集的向量形式是矩阵形式. 所以, 在我们的实验之前, 我们应该先格式化样本集. 样品的格式如下: <label> <value1> <value2>….这里在这里, 标签是网页类别, 值是特征权重. 预处理后, UCI 数据集转成到矢量格式. 然后, 我们使用这两种算法来做的学习. 实验的结果如下.

(1) 传统支持向量机算法的实验结果如表 3 所示.

表 3 传统 SVM 算法实验结果

训练集	训练集数	训练时间(秒)	少数类准确率(%)	少数类召回率(%)
Iris	99	17.8	74.5	87.6
Abalone	3300	443.6	71.3	82.3
Balance Scale	433	78.9	69.6	85.7
Yeast	1023	143.2	60.7	86.1

(2) 随机抽样 SVM 算法等结果如表 4 所示.

表 4 随机抽样 SVM 算法实验结果

训练集	训练集数	训练时间(秒)	少数类准确率(%)	少数类召回率(%)
Iris	99	17.8	87.7	92.5
Abalone	3300	443.6	81.9	87.4
Balance Scale	433	78.9	78.3	88.6
Yeast	1023	143.2	70.6	90.7

(3) Under-sampling-SVM 算法等结果如表 5 所示.

表 5 Under-sampling –SVM 实验结果

训练集	训练集数目	第一次训练时间(秒)	第二次训练时间(秒)	少数类准确率(%)	少数类召回率(%)
Iris	99	17.8	13.9	91.2	93.5
Abalone	3300	137.4	88.3	90.1	91.7
Balance Scale	433	18.9	15.6	88.5	93.3
Yeast	1023	21.7	16.9	85.6	94.8

从上表的实验结果, 我们可以看到, 从训练时间上 Under-sampling-SVM 比随机采样 VM 算法的时间较长, 但短于传统的支持向量机. 从少数类别的查准率和查全率来看, Under-sampling-SVM 算法具有更优秀的表现, 原因是经过欠采样, 改善了超平面的位置. Under-sampling-SVM 算法比随机采样 SVM 具有更好的表现原因是, Under-sampling-SVM 算法通过反馈机制改善了超平面位置.

5 不平衡SVM网页分类分类系统实现

5.1 实验环境

本文所设计的网页分类系统是在 visual studio 2010 编程工具基础上开发的, C++设计语言具备结构化控制语句, 程序的执行效率高, 而且 C++语言还具有汇编语言的优点, 并支持 C 语言.

5.2 不平衡 SVM 网页分类系统框架

本文设计的不平衡网页分类系统框架如图 2 所示.

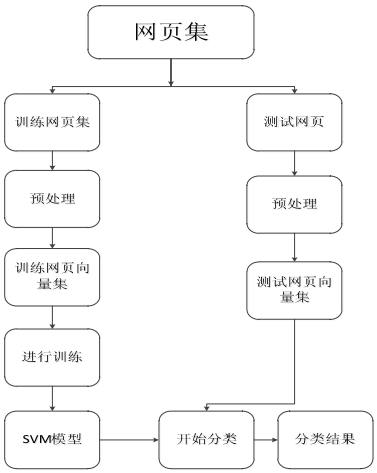

图 2 网页分类流程

5.2.1 网页预处理子系统

网页预处理主要包括 3 个步骤:

第一步: 网页去噪, 因为分类系统处理的对象主要是 web 页面, 但是页面中含有的元素丰富, 像文本, 图像, 音乐, 动态图片, 视频, 以及 html 标签, 这些因素会干扰我们处理信息. 经过大量的研究发现 HTML 的网页大多具有相同的格式, 如图 3 所示.

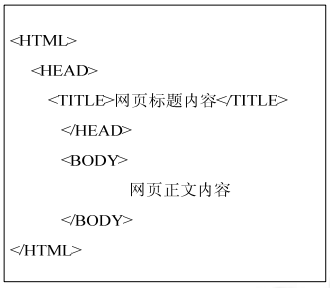

图 3　HTML DOM 树

所以本文的去噪方法主要是, 去掉网页中的“script”, “image”, “style”, “form”, “iframe”等标签.

第二步: 相关信息的提取主要是提取的 TD 标签, 表标签, div 标签的里的信息和相关链接的信息, 在这个时候只有去除对分类的贡献没有 HTML 标签, 保留中文字符即可.

第三步: 对提取好的有用信息, 进行中文分词, 本文采用的是中科院分词软件(ICTCLA).

5.2.2 预处理子系统

在经过对原始网页样本集进行净化之后, 这些网页就被变成了分类系统能识别的样本, 训练的过程用这测试数据得带基于欠采样不平衡 SVM 分类的初始模型, 不平衡 SVM 数据预处理子系统的过程如图 4 所示.

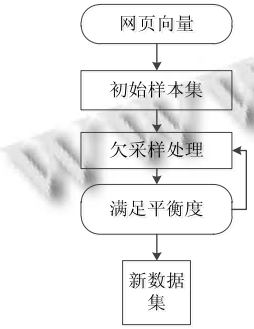

图 4　预处理子系统流程

5.2.1 分类子系统

在经过预处子系统处理过之后, 得到一个不平衡 SVM 模型, 接下来用设计好的分类子系统对网页进行分类判断, 主要的过程如图 5 所示.

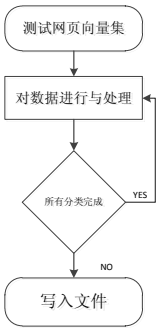

图 5　分类子系统流程

5.3 不平衡 SVM 网页分类系统界面以及运行结果

本文设计的系统的主界面图 6 所示.

图 6　系统的主界面

第一步. 先将本地的测试数据集读取到该系统中, 在点击界面中的“选择本地样本集”, 如图 7 所示.

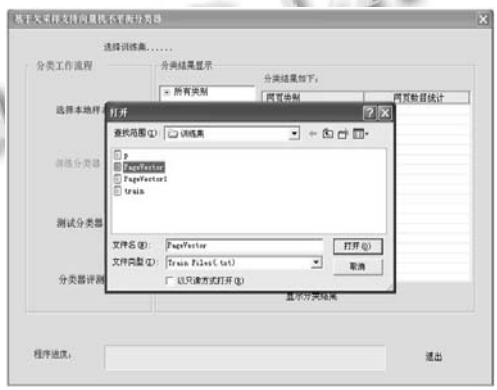

图 7　分类系统选取不平衡测试集界面

第二步. 按“训练分类器”按钮, 此时该分类系统变会对本地的样本集进行训练, 经过训练之后, 分类器系统会把样本训练集进行提取, 之后得到本地样本集的特征.

第三步. 训练结段结束后, 点击“测试分类器”按钮, 接下来, 该系统会显示分类效果的界面, 如图 8 所示.

图 8 系统分类结果图

该系统还有一个特点, 当你点击其中一个主题是时, 该体统会弹出一个界面, 上面显示着网页的具体信息, 像网页的大小、类别, 如图 9 所示.

图 9 系统的主界面

内存占用小、运行稳定, 分类精度高等优点, 具有较强的实用价值.

第四步: 接下来点击”分类器评测”按钮, 该系统评测界面, 该页面主要包括查准率和查全率这两大标准.

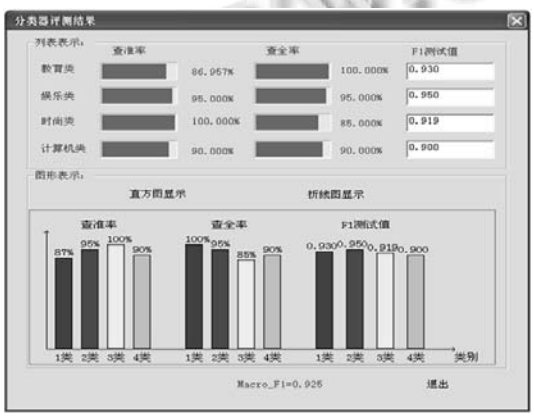

图 10 分类系统的评测结果

参考文献

1 Priore P, Parreno J, Pino R, et al. Learning-based scheduling of flexible manufacturing systems using support vector machines. Applied Artificial Intelligence, 2010, 24(3): 194–209

2 刘苏苏,丁福利,孙立民.优化支持向量机核参数的核矩阵方法研究.烟台大学学报(自然科学与工程版),2013, 26(2):131–135.

3 韩芳,孙立民.不平衡样本集分类算法研究.计算机应用研究, 2015,32(8):2323–2325.

4 杨智明.面向不平衡数据的支持向量机分类方法研究[博士学位论文].哈尔滨:哈尔滨工业大学,2009.

5 丁福利,孙立民.基于支持向量机的不平衡样本分类研究.科学技术与工程,2014,14(3):81–85.

6 刘苏苏,孙立民.支持向量机与 RBF 神经网络回归性能比较研究.计算机工程与设计,2012,32(12):4202–4205.

7 Vapnik VN. The nature of statistical learning theory. IEEE Trans. on Neural Networks, 1995, 8(6): 1564–1564.

8 李琼,陈利.一种改进的支持向量机文本分类方法.计算机技术与发展,2015,(5):78–82.

9 Joachims T. Making large-scale SVM learning practial. Advances in Kernel Methods-Support Vector Learning, MIT Press, 1999.

10 Platt JC. Sequential minimal optimization: A fast algorithm for training support vector machines. Advances in Kernel Methods-Support Vector Learning, 1998: 212–223.

11 柴岩,王庆菊.基于边界向量的样本取样 SMO 算法.系统工程,2015,(6):142–145.

12 Hwang JP, Park S, Kim E. A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. Expert Systems with Applications, 2011, 38: 8580–8585.

13 Liu Y, Yu XH, Huang JX, An A. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Information Processing and Management, 2011, 47: 617–631.

14 Dendamrongvit S, Kubat M. Undersampling approach for imbalanced training sets and induction from multi-label text-categorization domains. New Frontiers in Applied Data Mining, 2010, 5669: 40–52.

15 曾一平.中文文本情感分类的研究[硕士学位论文].北京:北京交通大学,2011.

16 余桂兰,陈珂,左敬龙.基于云模型的并行蚁群-SVM 分类方法.计算机技术与发展,2014,(4):131–134.

17 Frank A, Asuncion A. UCI machine learning repository. Irvine. CA: University of California, School of Information and Computer Science, 2010.