一种基于混合策略的失衡数据集分类方法

李 鹏,王晓龙,刘远超,王宝勋
(哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 提出了一种有效应用于失衡数据集的分类方法,其核心思想是从样本预处理和分类器改进两方面入手,为失衡数据集的分类问题提供全面的解决方案.首先创造性地采用动态自组织映射聚类的方法对失衡数据集进行重采样,这种采样方法,有效地解决了传统重采样的方法随机性强,人为主观干扰以及信息损失等弊端.随后借助 K-近邻规则的思想,对新采集的样本进行剪枝,有效地解决了实际存在的数据混叠现象.算法对 SVM 的核函数进行等角变换,由此对类边界进行了校准,以适应样本类别失衡的情况.通过对三种算法的对比实验证明了算法在失衡数据集分类上的有效性.本文的算法已经在答案抽取技术中得到了成功应用,并在 TREC2006 国际 QA 评测中得到了客观充分的验证.

关键词: 失衡数据集; 分类; 支持向量机; 动态自组织映射; K-近邻
中图分类号: TP391.2 **文献标识码:** A **文章编号:** 0372-2112 (2007) 11-2161-06

A Classification Method for Imbalance Data Set Based on Hybrid Strategy

LI Peng, WANG Xiao-long, LIU Yuan-chao, WANG Bao-xun
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: This paper presents a novel and effective classification method for imbalanced data sets. The core idea of the algorithm, which is composed of three parts, is to provide a general solution for IDS classification by both sample preprocessing and classifier improving. Firstly, we re-sample the imbalance data by using variable SOM clustering so as to overcome the flaws of the traditional re-sampling methods, such as serious randomness, subjective interference and information loss. Then we cut down the sampled data sets according to the K-NN rule to solve the problem of data confusion, which improves the generalization of SVM. Especially, in order to adapt the class imbalance, the class boundary alignment is introduced through conformal transform on kernel function. The comparison results show the effectiveness of three algorithms. Meanwhile, the algorithm has also been used in our question answer system, which obtains outstanding result in the international TREC-2006 QA track.

Key words: imbalanced data sets (IDS); classification; support vector machine (SVM); variable self-organizing maps (V-SOM); K-nearest neighbor (K-NN)

1 引言

目前的分类技术已经能较好地解决大部分具有数据量相对较小,标注比较完整及数据分布相对均匀等特点的问题和应用.但是,分类技术的发展与大规模应用仍受到很多问题的困扰,如:海量数据,数据集失衡,标注瓶颈以及数据混叠等.其中,失衡数据集的分类问题是分类技术研究中最具有挑战性的难点之一,得到广大研究人员的极大重视[1].

许多研究已经表明,对于失衡数据直接应用一些标准分类模型,如神经网络、支持向量机以及 C4.5 等,不能得到令人满意的分类效果[2].目前,解决数据失衡问题主要采取两种策略:一是重采样,可以适当屏蔽大类的信息量或提高小类的分类错误代价[3,4];二是采用新的分类策略,针对失衡数据的特点对分类算法进行改进以适应其特点[5~7].但是,目前所有方法在稀有类别上的分类准确性均很低,都不能将对稀有类别的识别水平整体提高到实际可以接受的程度,相关的研究仍需要进一步深入[8],研究人员面临着巨大的挑战.

本文的方法实现了样本重采样与分类器改进两种策略的有机结合.创造性地采用无监督聚类与 K-近邻规则相结合的重采样方法,对失衡数据集进行样本选取与剪枝,不但有效地平衡了数据偏斜状态,而且大大减少了支持向量的个数.此采样方法克服了传统采样方法存在的缺乏理论依据,随机性强,人为主观性干扰,信息损失等不足,同时对数据中存在的混叠现象予以很好的解决,显著提高了后续 SVM 分类器的泛化性能.为适应样本失衡的状态,我们对 SVM 分类模型也进行了改进.通过对核函数采用等角变换,以达到类边界校准的目的.值得指出的是,本文的方法实际应用于答案抽取技术中得到了令人满意的结果,在国际 TREC2006 QA 评测任务中得到了客观充分的评价.

收稿日期:2007-04-11;修回日期:2007-08-30
基金项目:国家自然科学基金重点项目(No.60435020);国家 863 高技术研究发展计划重点项目(No.2006AA01Z197)

2 失衡数据集采样对 SVM 分类的影响分析

失衡数据存在两个内在因素,即偏斜率与信息匮乏.偏斜率是指大类别与小类别的比值,它代表了数据失衡的程度.信息匮乏是指小类别样本的数据量,它表示了数据集中小类别的信息量.本文只讨论失衡数据集分类问题中应用最广泛的二分类问题,并且默认反例的数量级远大于正例.

支持向量机已经成功地应用于信息检索、图像识别以及文本分类等诸多领域[9,10].但是,当面对失衡数据集时,它的性能也显著下降[11].究其原因,主要是由于训练数据的失衡,使得正、反例支持向量的比率也明显失衡,反例起到主导地位而淹没正例,最终使得决策函数过多地将分类结果倾向于反例.

我们通过线性可分的失衡数据集来分析数据失衡以及采样对 SVM 分类的影响.从图 1(a)的训练过程可以看到,由于正、反例的失衡造成实际学习得到的分类超平面虽然在方向上与理想超平面基本保持一致,但远离反例而靠近正例,这是数据淹没现象造成的结果.如图 1(b)所示,这样的分类超平面在测试时会对反例有较强的倾向性,使一些正例被错分为反例.

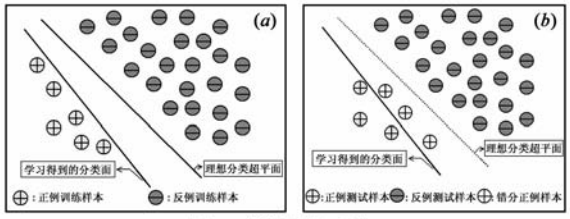

图 1 数据淹没现象

我们随机从反例中选取与正例相同数量的样本,使数据达到平衡状态.图 2(a)是重采样后的训练结果,虽然学习得到的分类超平面与正、反例之间的距离基本达到了理想状态,但与理想超平面的方向产生了很大的偏离,这是采样后信息损失造成的结果.如图 2(b)所示,这样的分类超平面在测试时也会有错分情况的发生.

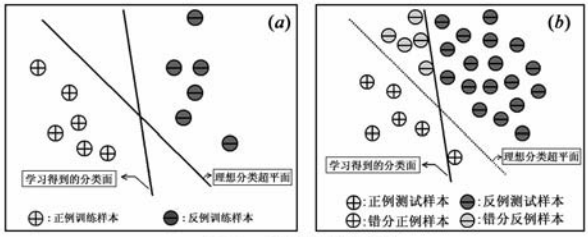

图 2 信息损失现象

因此,如何能够在降低偏斜率的同时,最大限度减少信息损失的发生是采用重采样方法解决失衡数据集分类所必需考虑的问题.

3 基于混合策略的失衡数据分类算法

本文的分类方法由三个子算法所组成,它们分别是基于动态自组织映射聚类的样本选取算法、基于 K-近邻规则的样本剪枝算法以及基于核函数变换的类边界校准算法.算法从样本重采样和 SVM 算法本身改进两方面入手,针对失衡数据集分类问题给出了比较全面的解决方案.

3.1 基于动态自组织映射聚类的样本选取算法

重采样的方法是解决数据失衡的一个有效途径,其关键是在于如何既能消除大量的噪声信息显著减小数据偏斜程度,又能保证最小的信息损失以保留绝大多数对分类学习有用的样本点.本文采用动态自组织映射聚类的方法来解决这个具有一定悖论性的难点问题.通过聚类的方法将原始的大规模失衡数据分为 N 个簇,并删除样本点均为反例的簇,将其它簇作为选取的样本集合.

采用动态自组织映射(V-SOM, Variable SOM)的聚类方法是为了避免由于神经元扩充可能导致的神经元欠利用现象,并且还可以克服矩形结构和其它结构容易带来的边界效应问题[12].神经元的权值调整采用如下公式:

$$\boldsymbol{n}_j(t+1)=\boldsymbol{n}_j(t)+\varphi(t)\cdot r_j(t)\cdot dis(x_i,\boldsymbol{n}_j(t)) \quad (1)$$

$$dis(x_i,\boldsymbol{n}_j(t))=1-sim(x_i,\boldsymbol{n}_j(t)) \quad (2)$$

其中 $\boldsymbol{n}_j(t+1)$和 $\boldsymbol{n}_j(t)$分别表示神经元 $\boldsymbol{n}_j$ 调整后和调整前的权值向量.$\varphi(t)$为学习速率函数,$r_j(t)$为邻域函数,二者随着训练的进行而逐渐递减.$dis(x_i,n_j(t))$表示样本向量 $\boldsymbol{x}_i$ 和神经元向量 $\boldsymbol{n}_j(t)$的距离,其大小可以转化为相似度的计算.向量之间的相似度越大,则其距离越小.一般地,其相似度可以采用余弦公式来计算,即

$$sim(\boldsymbol{x},\boldsymbol{n})=\frac{\sum_{i=1}^{l}W_{\boldsymbol{x}_i}W_{\boldsymbol{n}_i}}{\sqrt{\sum_{i=1}^{l}W_{\boldsymbol{x}_i}^2}\sqrt{\sum_{i=1}^{l}W_{\boldsymbol{n}_i}^2}} \quad (3)$$

公式(3)中 l 表示向量的维数,W_{x_i}表示样本向量 $\boldsymbol{x}$ 在第 i 维上的权值,W_{n_i}表示神经元向量 $\boldsymbol{n}$ 在第 i 维上的权值,文中涉及到的向量均经过归一化处理.

算法采用 R^2 聚类准则系数作为判断依据,在神经元的过利用和欠利用之间寻求平衡.令 $\boldsymbol{m}_i$ 为神经元 N_i 所对应的向量,则 N_i 所映射样本的类内离差平方和为

$$S_i=\sum_{x_j\to N_i}dis(x_j,\boldsymbol{m}_i) \quad (4)$$

在时刻 t,假设输出层共有 c 个神经元,则定义 $P_c=\sum_{k=1}^{c}S_k$.假设 T 为所有样本的总离差平方和,则 $T=$

$\sum_{i=1}^{|D|} dis(\boldsymbol{x}_i, \bar{\boldsymbol{x}})$

其中,$\bar{\boldsymbol{x}} = \frac{1}{|D|}\sum_{i=1}^{|D|}\boldsymbol{x}_i$ 表示所有训练样本的均值向量.$|D|$表示输入样本的总数,则

$$R^2 = 1 - (P_c / T) \tag{5}$$

聚类准则系数 R^2 的取值范围为$[0,1]$,并且其具体取值一般随着网络规模的增长呈单调增加的趋势.因此需要设定阈值 μ 以在适当的时候终止网络的增长,防止出现神经元的欠利用现象.

3.2　基于 K-近邻规则的样本剪枝算法

在真实的失衡数据集中,样本集合会存在数据混叠的情况.这种数据混叠在增大训练难度的同时还会造成过学习,使得 SVM 的泛化能力大大降低,分类性能明显下降[13].本文提出了基于 K-近邻规则的样本剪枝算法来解决这一实际应用中的难点.

算法的基本思想是对得到的新样本集合中的每个样本点考察与其最近的 K 个近邻样本的类别属性,通过计算当前查询样本点的预测值来判断查询样本的计算属性与本身真实属性是否一致.在实际应用中,正例的控制阈值通常小于反例的控制阈值,这是由于在失衡数据中正例资源本身就比较匮乏,正例信息相比反例信息比较珍贵,并且由于数据的失衡,混杂在正例中的反例远比混杂在反例中的正例多.因此,通过不同的控制阈值使修剪更倾向于删除反例混杂点,而保证稀有的正例信息尽可能不受损失.

算法假定所有的样本实例对应于 n 维空间中的点.更精确地讲,把任意的实例 $\boldsymbol{x}$ 表示为下面的特征向量,并采用标准欧氏距离作为两个向量之间的距离.

$$\langle \alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x}), \cdots, \alpha_n(\boldsymbol{x}) \rangle \tag{6}$$

其中,$\alpha_k(\boldsymbol{x})$表示实例 $\boldsymbol{x}$ 的第 k 个属性.那么两个实例 x_i 和 x_j 间的距离定义为

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{k=1}^{n}(\alpha_k(\boldsymbol{x}_i) - \alpha_k(\boldsymbol{x}_j))^2} \tag{7}$$

定义实例的类别属性值为 $f(\boldsymbol{x}_i) \in \{1, -1\}$,查询点的属性预测值 $\Psi(x_q)$由下面的公式计算得出:

$$\Psi(x_q) = \frac{\sum_{i=1}^{k} f(\boldsymbol{x}_i)}{k} \tag{8}$$

3.3　基于核函数变换的类边界校准算法

经过 V-SOM 聚类和 K 近邻剪枝处理后,失衡数据的失衡程度虽然被大大降低了,但数据集还仍然是一个失衡数据集.我们采用基于核函数等角变换的方法对类边界进行校准,以适应样本数据失衡的情况.

众所周知,核方法的基本思想是试图将输入空间 I 下线性不可分的样本通过 $\Phi(x)$映射到高维的希尔伯特特征空间 F 中,以得到高维空间下的线性分类面.在空间 F 中,所有的投影点将落在一个由 $\Phi(x)$所决定的曲面 S 上,在 $\Phi(x)$为连续可导的情况下,S 为一个扭曲的子流型,在其上可以计算两个点之间的黎曼距离,与欧几里德距离有所不同的是,黎曼距离是沿着曲面 S 上的路径做积分得到的.下面给出黎曼距离的计算式:

$$\mathrm{d}s^2 = \sum_{i,j} g_{ij}\mathrm{d}\boldsymbol{x}_i\mathrm{d}\boldsymbol{x}_j \tag{9}$$

其中因子 g_{ij}被称为黎曼度量,它与映射 $\Phi(x)$有关.尽管在核方法中 $\Phi(x)$并不是显式的给出的,但是我们总是可以通过核函数相关的计算技巧得到 g_{ij}与核函数 K 有关的表达式:

$$g_{ij}(\boldsymbol{x}) = \left(\frac{\partial^2 K(\boldsymbol{x}, \boldsymbol{x}')}{\partial \boldsymbol{x}_i \partial \boldsymbol{x}'_j}\right)_{X'=X} \tag{10}$$

显而易见,黎曼度量 $g_{ij}(\boldsymbol{x})$表征了输入空间 I 中的一个小区域在被映射 $\Phi(x)$映射到特征空间 F 以后被放大的程度,也正是这一点成为了核函数变换的基础.为了克服失衡数据给分类模型训练带来的干扰,我们试图通过对核函数 K 进行变换,从而扩大分类边界附近的黎曼度量 $g_{ij}(x)$,同时在其余的样本点上对其缩减.于是,我们引入核函数的等角变换式:

$$\tilde{K}(X, X') = D(X)D(X')K(X, X') \tag{11}$$

其中 $D(X)$为投影函数,我们必须对其进行适当的选择,才可以保证得到的 $\tilde{g}_{ij}(x)$在分类边界处具有较大的值,因此选择 RBF 距离函数的形式:

$$D(X) = \sum_{k \in SV} \exp\left(-\frac{|\boldsymbol{x} - \boldsymbol{x}_k|}{\tau_k^2}\right) \tag{12}$$

其中 τ_k^2 是一个重要的参数,它必须使 $D(X)$在距离分类界面较近的地方取值较大,而在远离分类界面的地方取值较小,因此可以按照下式设置:

$$\tau_k^2 = AVG_{i \in \{\|\Phi(X_i) - \Phi(X_k)\|^2 < M, y_i \neq y_k\}}(\|\Phi(X_i) - \Phi(X_k)\|^2) \tag{13}$$

其中 M 为一给定的距离常量.而 $\|\Phi(X_i) - \Phi(X_k)\|^2$ 可以通过核函数的运算得到:

$$\|\Phi(X_i) - \Phi(X_k)\|^2 = K(X_i, X_i) + K(X_k, X_k) - 2K(X_i, X_k) \tag{14}$$

为了处理分类边界处正例和反例的支持向量数量不平衡的情况,可以根据两类支持向量的数量对 τ_k^2 的系数进行调整.

4　实验

4.1　算法有效性验证与分析

为了验证本文三种算法对失衡数据集分类效果的作用与影响情况,本文在标准数据集 MUC-6 和 UCI 上选取了四个数据集进行验证,数据组成如表 1 所示.由于在失衡数据集分类的实际应用中,人们往往更关心

小类别的分类效果.因此,在本文的实验中选取正例分类的准确率(*Precision*)、召回率(*Recall*)以及 *F* 值(*F-measure*)作为衡量分类效果的技术指标.

表 1 失衡数据集列表

数据集	反例样本数	正例样本数	失衡比
MUC-6	159815	11266	14.2:1
UCI-Seg1	1980	330	6:1
UCI-Glass7	185	29	6.4:1
UCI-Abalone	4145	32	130:1

本文采用四种方法策略在上面四种不同的数据集上进行实验以验证三种算法在分类上所起的作用.实验中将每种数据集的 50% 用于训练,剩余的 50% 用于测试,并保证训练和测试数据具有相同的失衡比.四种方法的详细策略如下:

方法一:使用普通 RBF 核函数的 SVM 模型进行分类;方法二:使用基于核变换的 SVM 模型进行分类;方法三:在方法二的基础上引入 KNN 剪枝算法;方法四:在方法三的基础上引入聚类抽样算法.

通过对比方法一和方法二的实验结果可以验证核变换对分类效果的影响;通过对比方法二和方法三的实验结果可以验证 KNN 剪枝算法对分类效果的影响;通过对比方法三和方法四的实验结果可以验证聚类抽样对分类效果的影响.具体实验结果如图 3 所示.

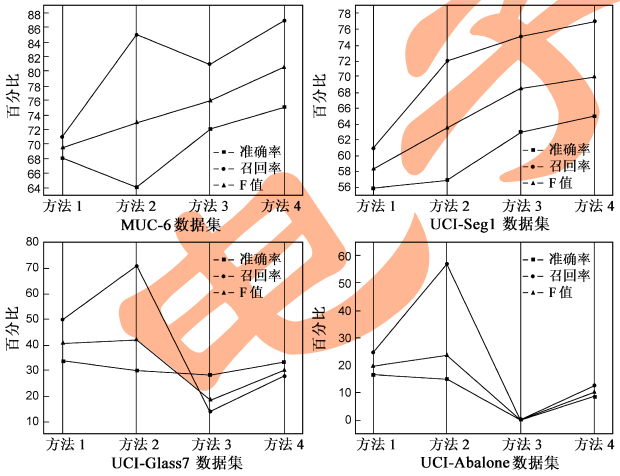

图 3 算法验证实验结果

通过对以上实验结果的对比与分析,我们可以得到一些很有意义的结论.(1)针对失衡数据分类,采用核变换的方法可以显著提高召回率,最终使系统分类的整体性能有所提高.也就是说通过采用核变换的方法可以找到更多的正例.(2)针对失衡数据分类,采用 KNN 剪枝算法可以提高分类的准确率,最终使系统分类的整体性能有所提高.也就是说通过采用 KNN 剪枝算法可以减少反例的错分情况.不过此算法在 UCI-Glass7 和 UCI-Abalone 两个数据集上的分类实验中造成系统性能严重下降,甚至失效.我们分析认为这是由于正例信息严重匮乏造成的结果,而与数据的失衡比的大小关系不大.因为 UCI-Glass7 和 UCI-Seg1 两个数据集的失衡比基本相同,但 UCI-Glass7 数据集中只有 29 个,训练数据中的正例只有 14 个,通过剪枝后所剩的正例信息更少,造成了分类效果的明显下降;在 UCI-Abalone 数据集中由于本来就极少的正例几乎被完全删除,造成了方法失效.因此,我们可以看到在失衡数据分类中不应仅考虑失衡比,信息匮乏程度也是影响分类方法和分类效果的重要因素.(3)针对失衡数据分类,采用动态 SOM 聚类抽样的方法对分类的准确率和召回率都有所提高,也进一步证明了重采样方法是解决失衡数据分类的一个有效途径.

4.2 实际参加 TREC QA 评测结果

文本检索会议(text retrieval conference,TREC)是由美国国家标准技术局(NIST)和国防部高级研究计划局(DARPA)组织召开的全球信息检索领域最具有权威性的评测会议.面向开放域的问答系统评测是会议的一个重要子任务,TREC 的 QA 评测已经被公认为问答系统全球规模最大,最具权威的评测公共平台,代表了开放域问答系统的最高水平.

答案抽取是问答系统的核心组成部分,也是信息抽取研究的一个子领域[14].对于一个基于实例的自然语言问题,往往只有一个正确答案,而相对应的干扰答案可能有十几,甚至几十个.因此,答案抽取是失衡数据集分类问题的一个典型应用领域.

我们的问答系统参加了 TREC 2006 QA 任务的评测,在全世界总共 59 个参评的系统中,我们的系统 InsunQA06 以准确率 29.8% 获得了 FACTOID 项目国际第五(国内第一)的成绩[15].同时,NIL 的准确率和召回率分别达到了 11.8% 和 35.3%,为本次国际评测的最好结果,NIL 的结果代表了系统处理复杂问题的能力,是问答系统智能程度的一个重要评价指标.本次参赛的问答系统中答案抽取部分的主体技术就是采用本文的算法,当然还有形式化答案抽取与网络信息抽取等技术相辅助.因此,可以说明本文的算法在答案抽取技术上的应用已经取得了效果.由此也可以证明本算法可以有效地解决失衡数据集分类问题中的实际问题,具有重要的现实意义.

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
lccPA06	Language Computer Corporation (Moldovan)	0.578	0.000	0.000
LCCFerret	Language Computer Corporation (Harabagiu)	0.538	–	0.000
cuhkqaepisto	The Chinese University of Hong Kong	0.390	0.107	0.353
ed06qar1	University of Edinburgh	0.323	0.069	0.294
InsunQA06	Harbin Institute of Technology (HIT)	0.298	0.118	0.353
QACTIS06A	National Security Agency (NSA)	0.266	0.118	0.118
ILQUA1	University of Albany	0.266	0.027	0.059
NUSCHUAQA1	National University of Singapore	0.261	0.000	0.000
asked06c	Tokyo Institute of Technology	0.251	–	0.000
QASCU3	Concordia University (Kosseim)	0.213	0.000	0.000

图 4 TREC 2006 QA 评测国际排名

5　结论

本文提出的失衡数据集分类算法,从失衡数据集样本的预处理和支持向量机核函数的改进两方面入手,为失衡数据集的分类问题提供了全面的解决方案.通过基于动态自组织映射聚类的样本选取算法,在最大限度地保存有效信息的同时对失衡数据中存在的大量噪声信息进行过滤,大大减少了支持向量个数,提高失衡数据分类精度和训练速度.这种采样方法,有效地解决了传统重采样的方法随机性强,人为主观干扰以及信息损失等弊端.通过基于 K-近邻规则的样本剪枝算法,对重采样的数据进行剪枝处理,用于解决实际存在的数据混叠问题,以提高 SVM 的泛化能力和分类的准确率.我们提出基于核函数变换的类边界校准算法,通过采用核函数等角变换的方法对类边界进行校准,以适应样本数据失衡的情况,实验结果表明其可以显著提高分类的召回率.在 MUC 和 UCI 四种标准数据集上的验证实验证明了本文的三种算法在失衡数据分类上的有效性.方法实际应用于国际 TREC2006 QA 评测比赛中取得了 29.8% 的准确率(59 个参赛系统中国际排名第五,国内第一),其中 NIL 的准确率和召回率分别达到了 11.8% 和 35.3%,为本次国际评测的最好结果.这表明本文的方法对处理失衡数据集分类问题具有一定程度的实用性.

参考文献:

[1] Chawla N V, et al. Editorial: special issue on learning from imbalanced data sets[J]. ACM SIGKDD Explorations, 2004, 6(1): 1 – 6.

[2] Batista G, et al. A study of the behavior of several methods for balancing machine learning[J]. ACM SIGKDD Explorations, 2004, 6(1): 20 – 29.

[3] Estabrooks A, et al. A multiple resampling method for learning from imbalanced data sets[J]. Computational Intelligence, 2004, 20(1): 18 – 36.

[4] Japkowicz N, et al. The class imbalance problem: a systematic study[J]. Intelligent Data Analysis, 2002, 6(5): 429 – 450.

[5] Japkowicz N, et al. Learning from imbalanced data sets: a comparison of various strategies[A]. Proceedings of the AAAI' 2000 Workshop on Imbalanced Data Sets[C]. CA: AAAI Press, 2000. 10 – 15.

[6] Provost F, et al. Machine learning from imbalanced data sets [A]. In Proceedings of the AAAI' 2000 Workshop on Imbalanced Data Sets[C]. CA: AAAI Press, 2000. 101 – 103.

[7] Visa S, et al. The effect of imbalanced data class distribution on fuzzy classifiers-experimental study[A]. In Proceedings of the FUZZ-IEEE Conference[C]. USA: IEEE Press, 2005. 22 – 26.

[8] 苏金树,张博锋,等.基于机器学习的文本分类技术研究进展[J].软件学报,2006,17(9):1848 – 1859.
Su Jinshu, Zhang Bofeng, et al. Advances in machine learning based text categorization[J]. Journal of Software, 2006, 17(9): 1848 – 1859. (in Chinese)

[9] 方景龙,陈铄,等.复杂分类问题支持向量机的简化[J].电子学报,2007,35(5):858 – 861.
Fang Jinglong, Chen Shuo, et al. A simplification to support vector machine for complicated recognition problem[J]. Acta Electronica Sinica, 2007, 35(5): 858 – 861. (in Chinese)

[10] 刘涵,郭勇,等.基于最小二乘支持向量机的图像边缘检测研究[J].电子学报,2007,34(7):1275 – 1279.
Liu Han, Guo Yong, et al. Edge detection based on least squares support vector machines[J]. Acta Electronica Sinica, 2007, 34(7): 1275 – 1279. (in Chinese)

[11] Akbani R, et al. Applying support vector machines to imbalanced datasets[A]. Proceedings of the 15th European Conference on Machine Learning[C]. Italy: Springer Press, 2004. 39 – 50.

[12] 刘远超.基于动态自组织映射模型的文本聚类研究[D].哈尔滨:哈尔滨工业大学,2006.

[13] 李红莲,袁保宗,等.一种改进的支持向量机 NN-SVM[J].计算机学报,2003,26(8):1015 – 1020.
Li Honglian, Yuan Baozong, et al. An improved SVM: NN-SVM[J]. Chinese Journal of Computer, 2003, 26(8): 1015 – 1020. (in Chinese)

[14] John O Connor, et al. Retrieval of answer sentences and answer-figures from papers by text searching[J]. Information Processing & Management, 1975, 20(1): 18 – 36.

[15] H Dang, J Lin, et al. Overview of the TREC 2006 question-answering track[A]. Proceedings of the text retrieval conference 2006[C]. USA. 2006. 1 – 16.

作者简介:

李　鹏　男,1978 年生于黑龙江哈尔滨.哈尔滨工业大学计算机科学与技术学院博士研究生.研究方向为问答系统、机器学习、网络信息处理.E-mail: pli@insun.hit.edu.cn

王晓龙　男,1955 年生于黑龙江哈尔滨.哈尔滨工业大学计算机科学与技术学院教授,博士生导师.研究方向为人工智能、机器学习、自然语言处理.