面向元信息分类的支持向量机改进技术

丁军平，蔡皖东
（西北工业大学计算机学院，710072，西安）

摘要：针对传统元信息分类方法的准确率不能满足主动P2P网络监测模型要求的问题，提出了一种基于改进支持向量机算法的元信息分类方法.该方法首先通过在加权最小二乘支持向量机的基础上加入对数据偏斜的处理，解决了元信息分类时关键词特征稀疏和样本高度不均衡问题，在对元信息文件名进行分词时，加入了词条之间的组合关系处理，在进行特征向量表示时，加入了对词条权值和语义属性的处理，最后使用基于粗糙集的属性规约方法进行特征向量选择，有效地降低了特征向量维度.实验结果表明，与传统方法相比，所提方法在进行元信息分类时能够大幅度提高分类准确率，准确率可达到97.8%，完全能够满足主动P2P网络监测模型的要求.

关键词：元信息分类；支持向量机；特征向量表示；粗糙集
中图分类号：TP393.0 **文献标志码**：A **文章编号**：0253-987X(2011)08-0037-06

An Improved Support Vector Machine Technology for Meta-Information Classification

DING Junping, CAI Wandong
(School of Computer Science and Engineering, Northwestern Polytechnical University, Xi′an 710072, China)

Abstract: An improved support vector machine(SVM) algorithm for meta-information classification is proposed to solve the problem that the traditional classification algorithm on meta-information can't meet the requirement of initiative P2P network monitoring model. The algorithm solves the problems that the keywords in classification are sparse and the distribution of sample is unbalanced by adding processing the data skew on LS-VSM. Combination relationships among keywords are added on filename segmentation of the meta-information. The weights of keywords and semantic attribute are processed for feature vector representation and the feature vector is selected using rough set specification so that the dimension of the feature vector is effectively reduced. Experimental results and comparisons with the traditional algorithm show that the proposed algorithm achieves a classification accuracy at 97.8%, that is the algorithm dramatically improves the classification accuracy, and meets the requirement of initiative P2P network monitoring model satisfactorily.

Keywords: meta-information classification; support vector machine; feature vector representation; rough set

传统的P2P网络流量识别及监控技术是一种整体识别技术，不能得到特定信息在P2P网络上的传输情况.在实际情况中，需要得到非法、有害文件的传输信息，对于正常文件的传输不需过多干预.在此需求下，作者提出了一种基于特定信息的主动监测模型，该模型的思路是：从网络上获取与特定信息

收稿日期：2010-12-10. 作者简介：丁军平(1977—)，男，博士生；蔡皖东(联系人)，男，教授，博士生导师. 基金项目：国家高技术研究发展计划资助项目(2009AA01Z424).
网络出版时间：2011-05-10 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20110510.0856.006.html

相关的元信息;对元信息进行解析,获取与文件相关的所有信息;根据已经获取的信息,使用伪客户端对特定信息进行监测,获取能够进行特定信息传输的节点列表信息;针对获取到的节点列表信息,逐个获得节点的详细状态信息.在此主动监测模型中,首先通过网络聚焦爬虫尽可能多地获取元信息,在获取的海量元信息中,如何找到需要被监控的非法元信息是进行后续操作的前提.

1 研究现状

元信息分类技术主要采用文本自动分类技术,文本自动分类是信息检索与数据挖掘领域的研究热点与核心技术[1],近年来得到了快速发展,被广泛应用于符号知识抽取、新闻分发、电子邮件排序以及用户兴趣学习等领域.文本自动分类的主要任务是在给定的分类体系下,根据文本的内容自动地确定与文本关联的类别.中文文本分类需要解决中文文本的表示方法、分词处理、文本的特征选择和特征抽取、选取分类模型和分类模型评估等问题.目前,文本分类技术以及相关的信息检索、信息抽取等领域已被广泛研究,产生了一些可用的分类系统,并取得了一定的成果.侯翠琴等[2]将网页之间的关联关系引入分类技术中,采用基于图的半监督学习算法对网页进行分类.现有的文本分类方法主要是针对长文本进行处理的,这些用于长文本的分类器方法都依据词频特征的相似性来分类.在元信息分类中,用于元信息分类的文件名长度比较短,一般来说不超过100个字符或者汉字.在进行分类时,它与普通文本主要有以下不同:元信息关键词特征稀疏.普通文本的特点就是文本内容比较长,而元信息中用于分类的文件名信息一般是短文本,由于元信息的关键词特征稀疏(每个短文本中一般只含有数十个甚至几个关键词),难以充分挖掘出特征之间的关联性,且样本高度不均衡.元信息分类时需要处理海量信息,分类后需要被真正监测的对象在数量上只占很小一部分,导致了样本和分类结果的高度不均衡性.传统分类器容易把大量的其他元信息错分为关注的监测对象.因此,还需要针对元信息样本高度不均衡的特点,对传统支持向量机(SVM)进行改造,以便得到结构更为合理的分类器来支持元信息的有效分类.

由于元信息分类和普通文本分类有以上的不同,使得传统的聚类分析方法在元信息表示这个层次上遇到了极大的困难[3],传统的文本表示模型都无法良好地表示.针对这种短文本的分类,国内的研究比较少,闫瑞[4]等使用将多种分类器动态组合的方法来完成对短文本的分类,取得了一定的效果,但由于采用多种分类器进行叠加,算法效率较差.樊兴华[5]等提出朴素贝叶斯(NB)算法与K近邻(K-NN)算法相结合的两步中文短文本分类方法,能被NB算法可靠分类的文本集直接使用NB分类,能被K-NN可靠分类的文本集直接使用K-NN分类,不能使用NB和K-NN可靠分类的文本集直接人工标注,这种方法由于事先不知道文本属于哪一个文本集,同一个文本最多会被判断3次,算法效率较差.宁亚辉[6]等提出了基于领域词语本体的短文本分类方法,首先抽取领域高频词作为特征词,借助知网从语义方面将特征词扩展为概念和义元,通过计算不同概念所包含相同义元的信息量来衡量词的相似度,从而进行分类.该方法在准确率方面有一定的提高,然而由于语言信息的复杂性和灵活性,该方法还在不断完善中.传统SVM方法是进行文本分类的经典方法,但其中核函数以及参数等的确定依据应用的不同千差万别.针对元信息分类的特点,本文对传统SVM进行改进,充分考虑数据偏移问题,在分词时考虑了词条之间的组合关系,在进行特征向量表示时考虑了词条的权值和语义属性,使得改进后的SVM算法能够提高元信息分类的准确率.

2 算法改进描述

2.1 改进的SVM描述

支持向量机是20世纪90年代最早由Vapnic[7]提出的一种通用学习算法,其数学描述如下:$\boldsymbol{x}$ 为 d 维空间中的向量,表示样本在特征属性上的离散归一值,y 为 $\boldsymbol{x}$ 的类标签.d 维空间中 n 个已经标记类别的训练向量记为$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,其中 $x_i\in\mathbf{R}^n$,$i=1,2,\cdots,n$($\mathbf{R}$ 为实数域).对于两分类问题,$y_i\in\{1,-1\}$,支持向量机分类问题可归结为下列规划问题

$$\left.\begin{aligned}&\min Q(\boldsymbol{w},\boldsymbol{\xi})=\frac{1}{2}\|\boldsymbol{w}^2\|+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2\\&\text{s.t:}\quad y_i=\boldsymbol{w}\varphi(x_i)+\boldsymbol{b}+\xi_i\\&\qquad\xi_i\geqslant 0,\ i=1,2,\cdots,n\end{aligned}\right\}\tag{1}$$

式中:C 为预先设定的惩罚系数,表示对干扰样本的容忍程度;ξ_i 为松弛变量.用拉格朗日法求解这个优化问题

$$L(\boldsymbol{w},\boldsymbol{b},\boldsymbol{\xi},\boldsymbol{a})=\frac{1}{2}\|\boldsymbol{w}^2\|+$$

$$\frac{C}{2}\sum_{i=1}^{n}\xi_i^2-\sum_{i=1}^{n}a_i[\boldsymbol{w}\varphi(x_i)+\boldsymbol{b}+\xi_i-y_i] \quad (2)$$

式中：$\boldsymbol{a}=[a_1,a_2,\cdots,a_n]^{\mathrm{T}}$ 为拉格朗日乘子. 由式(2)的平衡条件可知

$$\left.\begin{aligned}\frac{\partial L}{\partial \boldsymbol{w}}&=\boldsymbol{w}-\sum_{i=1}^{n}a_i\varphi(x_i)=0\\ \frac{\partial L}{\partial \boldsymbol{b}}&=-\sum_{i=1}^{n}a_i=0\\ \frac{\partial L}{\partial \xi_i}&=C\xi_i-a_i=0\\ \frac{\partial L}{\partial a_i}&=\boldsymbol{w}\varphi(x_i)+\boldsymbol{b}\,\xi_i-y_i=0\end{aligned}\right\} \quad (3)$$

可推导出

$$\begin{bmatrix}\boldsymbol{I}&\boldsymbol{0}&\boldsymbol{0}&-\boldsymbol{Z}\\ \boldsymbol{0}&\boldsymbol{0}&\boldsymbol{0}&-\bar{\boldsymbol{1}}\\ \boldsymbol{0}&\boldsymbol{0}&C\boldsymbol{I}&-\boldsymbol{I}\\ \boldsymbol{Z}&\bar{\boldsymbol{1}}&\boldsymbol{I}&\boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{w}\\ \boldsymbol{b}\\ \boldsymbol{\xi}\\ \boldsymbol{a}\end{bmatrix}=\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\\ \boldsymbol{0}\\ \boldsymbol{y}\end{bmatrix} \quad (4)$$

式中：$\boldsymbol{Z}=[\varphi(x_1),\varphi(x_2),\cdots,\varphi(x_n)]^{\mathrm{T}}$；$\boldsymbol{y}=[y_1,y_2,\cdots,y_n]^{\mathrm{T}}$；$\bar{\boldsymbol{1}}=[1,1,\cdots,1]^{\mathrm{T}}\in\mathbf{R}^n$；$\boldsymbol{\xi}=[\xi_1,\xi_2,\cdots,\xi_n]^{\mathrm{T}}$；$\boldsymbol{a}=[a_1,a_2,\cdots,a_n]^{\mathrm{T}}$. 由式(3)可知，$\boldsymbol{w}=\sum_{i=1}^{n}a_i\varphi(x_i)$，$\xi_i=\frac{a_i}{C}$，消去 $\boldsymbol{w}$ 和 ξ_i 并对线性方程组求解，可得

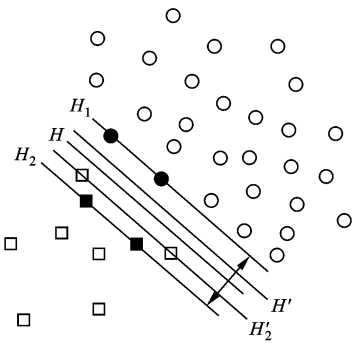
$$\left.\begin{aligned}\boldsymbol{a}&=\boldsymbol{A}^{-1}(\boldsymbol{y}-\boldsymbol{b}\bar{\boldsymbol{1}})\\ \boldsymbol{b}&=\frac{\bar{\boldsymbol{1}}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{y}}{\bar{\boldsymbol{1}}^{\mathrm{T}}\boldsymbol{A}^{-1}\bar{\boldsymbol{1}}}\end{aligned}\right\} \quad (5)$$

由式(5)和 $\boldsymbol{w}=\sum_{i=1}^{n}a_i\varphi(x_i)$，可得非线性预测模型

$$\boldsymbol{y}=\varphi(\boldsymbol{x})+\boldsymbol{b}=\sum_{i=1}^{n}a_i\varphi(x_i)\varphi(\boldsymbol{x})+\boldsymbol{b}=\sum_{i=1}^{n}a_iK(x_i,\boldsymbol{x})+\boldsymbol{b} \quad (6)$$

式中：$K(x_i,\boldsymbol{x})=\varphi(x_i)\varphi(\boldsymbol{x})$称为核函数. $\varphi(\boldsymbol{x})$的主要功能是接受两个低维空间里的向量，能够计算出经过某个变换后在高维空间里的向量内积值，是任何满足 Mercer 条件的函数. 核函数的种类很多，常用的有多项式核函数、Gaussian 径向基 RBF 核函数、线性核函数、Sigmoid 核函数等，选择不同类型的核函数，在学习能力和推广能力上会有所差异. 因为多项式核函数具有良好的全局性质、很强的外推能力，其阶数越低，推广能力越强，在传统文本分类方面已经有很好的表现. 元信息分类是基于短文本的文本分类技术，向量维数较低，关键词特征矩阵稀疏，使用多项式核函数能够充分发挥多项式核函数的优点，增强分类算法的推广能力，所以本文选择多项式核函数.

根据前面的分析可知，元信息的样本信息具有高度不均衡性，需要监测的样本只占所有样本的小部分，这种情况被称为数据偏斜问题，它会引起分类问题中超平面的偏移，如图 1 所示.



H 为中心分类面；H_1、H_2 为正常样本类和被监测样本类的分类面；H'、H'_2为增加 2 个被监测样本后 H 和 H_2 的新位置

图 1　样本分布示意图

本文解决数据偏斜问题的方法就是对不同的分类给予不同的惩罚因子 C，正常样本的数量较多，给予的 C 就小，被监测样本的数量较少，每个样本的重要性就高，给予的惩罚因子就大. 因此，目标函数中因松弛变量而损失的部分就变为

$$\frac{C_1}{2}\sum_{k=1}^{p}\xi_k^2+\frac{C_2}{2}\sum_{j=p+1}^{n}\xi_j^2 \quad (7)$$

式中：p 为正常样本的数量；$n-p$ 为被监测样本的数量；C_1 为正常样本的惩罚因子；C_2 为被监测样本的惩罚因子. C_1 与 C_2 之间的比例为样本在超维空间所占据的超球直径之反比. 样本超球直径的计算方法为首先计算本分类内所有样本之间的距离，再求最大距离，将最大距离作为超球直径即可. 由于 $n=p+q$，在元信息分类中，p 的数量要远大于 q，C_2 要远大于 C_1，C_1 和 C_2 根据正常样本和监测样本的特征向量是可计算的，ξ_k^2 与传统 SVM 中的 ξ_i^2 相同，所以式(7)是可计算的. 式(7)表示将原来所有样本因松弛变量而损失的部分变化为监测样本的损失与正常样本的损失之和，同时给予监测样本和正常样本不同的惩罚因子，即重要性度量标准，这种计算方式在实际计算中被证明是可行的.

在求解原始非线性预测模型时，可知松弛变量 $\xi_i=a_i/C$，根据最新的惩罚因子可知正常样本的松弛变量 $\xi_k=a_k/C_1$，$k=1,2,\cdots,p$，被监测样本的松

弛变量 $\xi_j=a_j/C_2$，$j=p+1,\cdots,n$，将 ξ_k、ξ_j 代入线性方程组(3)中求得 $\boldsymbol{a}$ 和 $\boldsymbol{b}$，从而得到最新的非线性预测模型.

2.2 基于互信息的关键词组合分词法

分词就是将连续的字符序列按照一定的规范重新组合成词条序列的过程. 当前自动分词的方法较多且相对成熟，常用的分词算法包括机械分词法、语义分词法和人工智能法.

本文的研究对象为元信息分类，元信息文件名长度一般不超过 100 个字符或者汉字，只有少量甚至没有标点符号出现，不会有过于复杂的语法结构出现，每个词条的出现次数基本上都是一次. 研究表明，双向最大匹配算法的准确度可以达到 95%以上，完全可以满足元信息分类对分词的要求，所以在本文中采用双向最大匹配法进行分词.

在元信息分词中，通过分词算法可以将元信息文件名准确地分为词条信息，但是在进行实际分词过程中发现，非法信息的文件名称为了防止被监测，在某些词条中间会加入干扰信息，例如词条“法轮功”会变为“法＊轮＊功”等等，对于这种情况，分词算法只能得到简单的法、轮、功 3 个字，而不能得到准确的分词信息. 针对此问题，我们对分词中的互信息概念进行扩充，提出词条互信息的概念.

定义 1 词条互信息 对词条序列 $A_1,A_2,\cdots,A_n$，定义它们之间的互信息为

$$I(A_1,A_2,\cdots,A_n)=\frac{P(A_1,A_2,\cdots,A_n)}{P(A_1),P(A_2),\cdots,P(A_n)}$$

式中：$P(A_1,A_2,\cdots,A_n)$表示词条 $A_1,A_2,\cdots,A_n$ 出现的概率；$P(A_1),P(A_2),\cdots,P(A_n)$分别代表 $A_1,A_2,\cdots,A_n$ 作为单个词条出现的概率. 这样定义的词条互信息能够反映出词条之间结合关系的紧密程度.

2.3 分词词典设计

分词词典的建立是进行分词的关键，传统的分词词典的元素结构一般定义为

(〈词条〉,〈词性〉)

分词词典由多个元素按照词条的顺序排列组合而成. 在这种结构定义中，元素中的〈词条〉信息是使用双向最大匹配算法进行分词的依据，〈词性〉表示该词条的词性，对分类结果起主要作用的是名词、动词和连词. 根据元信息分类的特点，本文对词典元素进行改进，使之能够更有效地进行元信息分类. 改进后的词典元素定义为

(〈词条〉,〈词性〉,〈语义属性〉,
〈权值 1〉,…,〈权值 n〉)

其中〈权值 i〉表示该词条属于分类 i 的概率，第 k 个词条的权值计算方法为

$$D_{ki}=F_{ki}\Big/\sum_{j=1}^{n}F_{kj} \tag{8}$$

式中：F_{ki}表示在分类 i 中出现第 k 个词条的样本个数. 本文所研究的元信息分类结果只有不被监测类和被监测类，所以只存在〈权值 1〉和〈权值 2〉.

〈语义属性〉表示该词条在语义方面的属性，主要用于处理连词中的语义转折词的属性，像“但是”、“不”、“不得”等连词，会对分类结果产生完全相反的影响，需要对它们的语义属性进行标注.〈语义属性〉的取值范围为{－1,1}，－1 表示需要进行语义转义，1 表示不需要进行语义转义.

2.4 基于权值和语义属性的特征向量表示方法

经过分词处理后的文档，仍不能直接被分类器解释和处理，为使计算机能够真正处理文本特征，必须将词作为向量的维数表示文本，将文本映射为计算机可以处理的数学向量. 常用的映射模型为向量空间模型，该模型能较为全面地反映文档的特性，模型中每一个元信息的文件名都被映射为一组规范化正交词条向量，表示为

$$\boldsymbol{T}(f)=(t_1,w_1;t_2,w_2;\cdots;t_i,w_i;\cdots;t_n,w_n) \tag{9}$$

式中：t_i 表示词条；w_i 表示词条 t_i 在元信息文件名 f 中对应的权重，统一的表示形式为 $w(t_i,f)$. 一般采用 TF-IDF 方法计算 $w(t_i,f)$的数值，但是传统的 TF-IDF 算法仅考虑了词语在文档集中的分布情况，而未考虑具体分布的比例. 为了更好地进行特征项的选择和提取，本文对 $w(t_i,f)$的计算式进行改进，改进后的计算式为

$$w(t_i,f)=\frac{\left|D_{i1}-\sum_{l=2}^{L}D_{il}\right|s(t_i,f)\mathrm{lb}(N/n_{ti}+r)}{\left(\sum_{t_j\in f}\left[\left|D_{j1}-\sum_{l=2}^{L}D_{jl}\right|s(t_j,f)\mathrm{lb}(N/n_{tj}+r)\right]^2\right)^{1/2}}Y_i \tag{10}$$

式中：$s(t_i,f)$表示词条 t_i 在 f 中出现的次数；t_j 表示在 f 中出现的所有词条；N 表示训练集中样本的总数；n_{ti}表示出现词条 t_i 的所有样本数；r 是常数，在实际使用中一般取 0.01；Y_i 表示词条 t_i 的语义属性；D_{il}为词条 t_i 的〈权值 l〉；L 表示分类的个数，在元信息分类中，$L=2$. 该公式在原有计算公式的基础上考虑了词条权值和语义属性信息，使得分类效

果更好.

2.5　特征向量选择

因自然语言文本集中往往包含大量的词汇，如果把这些词都作为特征，将使得文本向量空间的特征维数高达几万至几十万，分类器的算法和实现的复杂度都随着特征空间维数的增加而增加. 为了能有效地削减特征空间维度，解决的办法是降低特征向量的维度，选择那些最有代表意义的词条作为特征词条. 本文在已经计算出来的特征向量基础上，采用基于粗糙集的属性规约方法进行特征向量的选取.

粗糙集理论模型可以定义为 $K=\langle U,Q,V,e\rangle$，其中：$U=\{d_1,d_2,\cdots,d_N\}$ 为所有训练样本构成的论域；$Q=\{C,D\}$ 为属性集合，$C=\{t_1,t_2,\cdots,t_M\}$ 为条件属性，M 为特征向量的维数，$D=\{G_1,G_2,\cdots,G_L\}$ 为结果属性，表示元信息所属的分类；V 表示属性值的集合；e 为信息函数. 对 $\forall d\in U,q\in Q$，有 $e(d,q)=V_q$，V_q 表示属性 q 的属性值. 基于粗糙集的属性规约方法步骤描述如下：

(1)对样本的特征向量表示进行离散化；

(2)创建决策表 $K=\langle U,Q,V,e\rangle$；

(3)构造属性核 $L(C)$；

(4)初始化最小属性集 $\min p=\phi$；

(5)首先将属性核加入到最小属性集 $\min p$ 中，去掉属性核所在的列和在此属性上值不为 0 的所有行；

(6)如果决策表不为空，则转到步骤(7)，否则停止，此时 $\min p$ 即为一个最小属性约简集，也就是特征向量的选择结果；

(7)对剩余的决策表重新计算各列中值不为 0 的个数，选取个数最多的属性加入到 $\min p$ 中，去掉加入到 $\min p$ 中的属性所在的列以及该列中属性值不为 0 的元素对应的所有行，转到步骤(6).

3　实验验证

为了验证本文所描述算法的效率和准确性，首先通过网络爬虫程序从网络上获取大量的元信息，元信息包括 BT 种子文件和 Ed2k 链接，其中元信息总数为 27 123，BT 种子文件数为 17 668，Ed2k 链接数为 9 455. 由于我们主要根据元信息文件名进行分类，所以首先需要提取元信息中所包含的文件名信息. 得到文件名信息后，为了对算法进行训练，再使用其中 5 000 个元信息作为样本信息，其他元信息作为验证数据. 然后，对样本数据的分类进行人工标注，标注结果是需要被监测的元信息为 237 个，不需要被监测的元信息为 4 763 个.

虽然本文所涉及到的改进算法较多，主要包括对传统 SVM 的改进、对分词算法的改进、对特征向量表示的改进等等，但是所有的改进算法都是为提高 SVM 算法的分类效果服务的，所以本文主要针对传统 SVM 算法与改进后 SVM 算法的效率及准确率进行比较，涉及的内容包括样本学习的效率、分类的效率以及准确率.

对样本进行学习所花时间实验设计如下. 使用不同的样本集对传统 SVM 算法和改进后的 SVM 算法进行训练，其学习时间对比如图 2 所示.

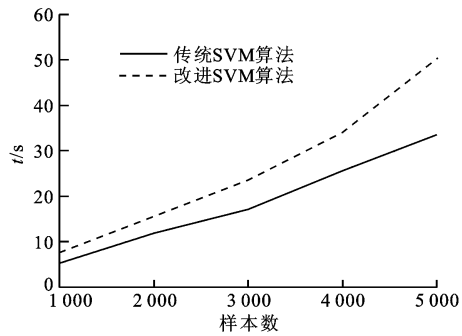


图 2　传统 SVM 算法与改进 SVM 算法的样本学习时间对比

当使用不同数量等级的样本集进行训练时，在相同的分词算法、特征向量表示和特征向量选择算法情况下，由于改进 SVM 算法增加了对数据偏斜的处理，故在对样本的学习速度方面明显低于传统的 SVM 算法.

由于我们的目的是进行元信息分类，因此看中的是这两种算法在实际分类过程中的效率和准确度. 使用不同数量的验证数据进行分类，两种算法的分类时间对比如图 3 所示，分类准确率对比如图 4 所示.

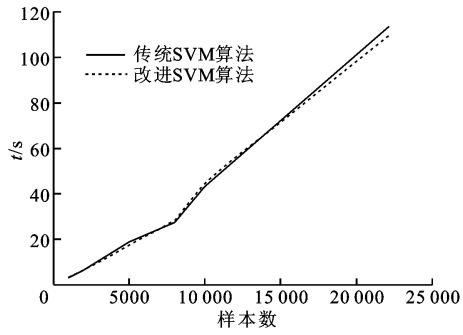


图 3　传统 SVM 算法与改进 SVM 算法的分类时间对比

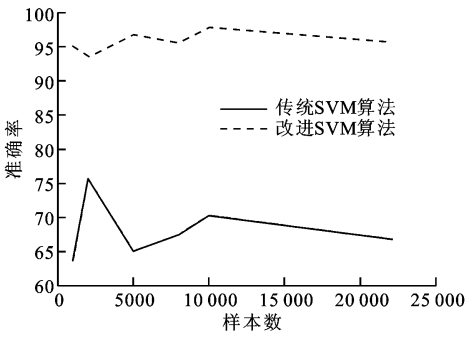


图4 传统SVM算法与改进SVM算法的分类准确率对比

由图3、4可知，当使用不同数量的验证数据进行元信息分类时，传统的SVM算法与改进SVM算法在分类时间上的差别不大，因为它们都是使用相同的公式进行计算，只是参数不同而已，但是在分类准确率方面，由于传统的SVM算法没有考虑数据偏移问题，导致分类结果准确率偏低，而改进SVM算法能够大幅度提高元信息分类的准确率.

通过上述实验验证可以看出，由于改进SVM算法增加了对数据偏移的处理，虽然在学习阶段增加了学习时间，降低了学习效率，但是在进行元信息分类时，能够在损失少量效率的情况下(最多损失6.5%)大幅度提高准确率，提高幅度可达49.5%，准确率可达到97.8%. 本文的实验结果表明，对SVM算法的改进是正确的，达到了预期的效果.

4 结 论

本文在进行元信息分类时，在加权最小二乘支持向量机的基础上加入了对数据偏斜的处理，极大地改善了元信息分类样本高度不均衡的特点；在进行元信息文件名分词时，加入了对词条之间组合关系的考虑，提高了对干扰信息的过滤；在进行特征向量表示时，加入了对词条权值、语义属性和统计信息的处理，并对异常数据进行了过滤，增加了特征向量的表述信息量；为了能有效地削减特征空间维度，使用基于粗糙集的属性规约方法进行特征向量的选择. 实验验证表明，改进后的SVM算法虽然在学习阶段所花的时间较多，但是能够在不损失分类效率的前提下大幅提高元信息分类的准确率.

参考文献：

[1] 徐沛娟，李雄飞，惠玥，等. 中文文本分类相关算法的研究与实现[J]. 吉林大学学报，2009，47(4)：790-794.
XU Peijuan，LI Xiongfei，HUI Yue，et al. Research and implementation of related algorithm of Chinese text categorization[J]. Journal of Jilin University，2009，47(4)：790-794.

[2] 侯翠琴，焦李成. 基于图的Co-Training网页分类[J]. 电子学报，2009，37(10)：2173-2219.
HOU Cuiqin，JIAO Licheng. Graph based Co-Training algorithm for web page classification[J]. Acta Electronica Sinica，2009，37(10)：2173-2219.

[3] 杨震，范科峰，雷建军，等. 基于语义的文本流形研究[J]. 电子学报，2009，37(3)：557-561.
YANG Zhen，FAN Kefeng，LEI Jianjun，et al. Text manifold based on semantic analysis[J]. Acta Electronica Sinica，2009，37(3)：557-561.

[4] 闫瑞，曹先彬，李凯. 面向短文本的动态组合分类算法[J]. 电子学报，2009，37(5)：1019-1024.
YAN Rui，CAO Xianbin，LI Kai. Dynamic assembly classification algorithm for short text[J]. Acta Electronica Sinica，2009，37(5)：1019-1024.

[5] 樊兴华，王鹏. 基于两步策略的中文短文本分类研究[J]. 大连海事大学学报，2008，34(3)：121-124.
FAN Xinghua，WANG Peng. Chinese short-text classification in two-steps[J]. Journal of Dalian Maritime University，2008，34(3)：121-124.

[6] 宁亚辉，樊兴华，吴渝. 基于领域词语本体的短文本分类[J]. 计算机科学，2009，36(3)：142-145.
NING Yahui，FAN Xinghua，WU Yu. Short text classification based on domain word ontology[J]. Computer Science，2009，36(3)：142-145.

[7] VAPNIC V. The nature of statistical learning theory[M]. New York，USA：Springer，1995.

(编辑 荆树蓉 武红江)