

商空间框架下的大规模 SVM 数据集约减法

覃 希 苏一丹 张 雯

(广西大学计算机与电子信息学院 南宁 530004)

摘 要 借助商空间框架下的粒度分析理论及其计算方法,提出将“粒度”的概念用于大规模 SVM 数据集的约减来建立商空间框架下的约减模型。该约减模型的约减方向是由远及近地向分类超平面削减,其削减幅度也伴随集合的缩小而由粗到细逐渐变化。同时,给出该模型的一种实现。实验证明,商空间框架下的 SVM 约减模型比普通 SVM 约减模型的压缩效果更好。

关键词 商空间,粒度,约减法,支持向量机

中图法分类号 TP391 **文献标识码** A

Reduction for Large-scale SVM Datasets under Quotient Space

QIN Xi SU Yi-dan ZHANG Wen

(School of Computer and Electronics Information, Guangxi University, Nanning 530004, China)

Abstract Using granularity analysis theory and computational method in quotient space, we are able to build a reduction model in quotient space. In this model, we cut out the redundant data using variable granularity so that the reduction becomes more accurate. An example implementation of this method was provided. Experiments indicate our new method yields significantly improved compression without sacrificing the accuracy of traditional SVM techniques.

Keywords Quotient space, Granularity, Reduction, SVM

1 引言

支持向量机(Support Vector Machines, SVM)是一种基于小样本统计学习的机器学习系统。在处理大规模问题时, SVM 会出现运算瓶颈。为此,对大样本问题做训练前的预处理是非常有必要的。文献[1]中提出的 TransRed 和 GetBorder 方法,能够确定支持向量的近似集合;文献[2]利用线性规划找到支持向量的一个超集;文献[3]将聚类应用于大样本集的预处理中,提出了 CB-SVM(Clustering-Based SVM);文献[4]将 K 均值、K-means 等聚类算法运用于大样本集的预处理;文献[5]利用核空间距离聚类法对训练样本进行缩减;文献[6]提出了两个新的聚类算法来约减 SVM 样本集;文献[7,8]则基于近邻原则来压缩样本点。以上的方法在约减样本数据集时都没有考虑“粒度”的概念,本文利用商空间框架思想,建立商空间框架下的约减模型,在约减法中体现“粒度”的概念,同时给出了该模型的一种实现方式——双层减样法(BR)。然后用基于商空间框架的减样法优化基于 SVM 的垃圾标签检测模型。

2 商空间模型

2.1 商空间框架下的机器学习方法

人类智能的其中一个特点是:对于同一个问题,人们既可

以从高处统观全局,进行整体思考,也能进入问题内部展开细节分析。这种全局与局部相结合的分析能力,是人类智能复杂性的一个体现。参照人类智能的这一特点,在人工智能中,也想体现这样的全局分析与局部分析相结合的能力。为此,研究人员利用相关的数学模型对其建模,并产生了新的机器学习方法——商空间框架下的机器学习法^[9,10],其中包含全局分析能力数学模型、局部分析能力数学模型和全局局部混合数学模型。

2.2 多粒度计算的商空间模型

全局分析能力的数学模型涉及“粒度”概念。在此首先给出关于粒度的定义^[10,12]。

定义 1(粒度的定义) 给定论域 U 和 U 上的一个关系 $R:U$,使得 $U=\bigcup_{i\in\tau}G_i$,那么称 G_i 为信息粒子, $\{G_i\}_{i\in\tau}$ 是论域的一种粒度。 R 可以代表等价关系、不可分关系、相似、相等、属性、投影、结构和一般函数等。

在定义 1 的基础上进一步讨论商空间下粒度分析的数学模型;对不同粒度的分析,就是给定一个等价关系,在这个等价关系下,将原空间中的问题对象变换成一个商空间中的对象,然后再对它进行分析。由此可以得到多粒度计算的商空间模型定义。

定义 2(多粒度计算的商空间模型定义) 给定一个研究的对象,论域记为 X ,在 X 上每个元素的属性记为 $f(f:X\rightarrow Y$,

到稿日期:2013-05-21 返修日期:2013-07-16 本文受国家自然科学基金(61063032),中国教育部人文社会科学研究规划基金(11YJAZH080)资助。

覃 希(1983—),女,硕士,助教,主要研究领域为 Web 标签挖掘,E-mail:suancai1412@163.com;苏一丹(1962—),男,博士,教授,硕士生导师,主要研究领域为电子政务、Web 个性化挖掘;张 雯(1985—),女,硕士生,主要研究领域为 Web 数据挖掘。

$f=(f_1 \cdots f_n)$ 。然后记 T 为 X 中各元素之间的关系,我们用 X 上的拓扑 T 来描述。这样就用三元组 (X, f, T) 来描述我们研究的对象。现在从某个粒度对 (X, f, T) 进行分析,即给定 X 上的一个等价关系 R ,记对应于 R 的商集为 $[X]$,由 f, T 将产生对应的 $[f], [T]$,称为商属性和商结构。这样我们就得到一个新的研究对象,记为 $([X], [f], [T])$,则 $([X], [f], [T])$ 称为 (X, f, T) 的商空间。

不同种粒度之间的关系可以用粒度的粗细来描述。

定义 3(粒度的粗细关系定义) 设 R 是论域上关系的全体,对于 $\forall x, y \in U, xR_1y \Rightarrow xR_2y, R_1, R_2 \in R$,则称 R_1 比 R_2 细,记为 $R_2 < R_1$ 。若一个关系代表一种分类,设 $R_0 < R_1 < \cdots < R_{end}$,表示一个嵌套关系簇, R_0 代表论域本身是一个等价类,且是最粗的划分; R_{end} 代表 $\forall x, y \in U, xR_{end}y \Leftrightarrow x=y$,这是最细的划分;其他表示中间层次的划分。

由定义 2、定义 3 得知,原问题空间的商空间是利用等价关系搭建的。通过建立不同的等价关系,就可以生成不同粒度的商集以及其上的商属性和商结构。特殊地,我们可以在商集上继续使用等价关系,建立商集上的商集。这样可以使商空间中的粒度放大。对于以往只能在某个固定的集合上完成的操作,可以通过调整对象间的等价关系,在一组嵌套关系簇上完成。

2.3 商空间中对论域取粒度的方法

对论域取粒度一般有投影和合成两种方法。以下是对投影和合成的定义^[13]:

定义 4(投影的定义) 设问题为 (X, f, T) ,给定 X 上的一个等价关系 R ,求其对应的商空间 $([X], [f], [T])$,即作 $[X]=\{[x]|x \in X\}, [x]=\{y|yRx, y \in [X]\}, [T]=\{u|p^{-1}(u) \in T, u \subset [X]\}$,其中 $p: X \rightarrow [X]$ 称自然投影, p^{-1} 是 p 的逆映射, $[f]$ 由 f 按一定原则构造。

定义 5(合成的定义) 已知若干个商空间 $(X_1, f_1, T_1), \cdots, (X_n, f_n, T_n)$,将它们合并成新的商空间 (X', f', T') ,其中 X' 是 $X_1, \cdots, X_n$ 的上确界空间, T' 是 $T_1, \cdots, T_n$ 的上确界拓扑, f' 满足 $p_i f' = f_i, p_i$ 是 X 到 X_i 的投影, $i=1, \cdots, n$,同时满足某个最优准则。

3 商空间框架下的约减模型

商空间理论向我们透露了“粒度”的概念。本文从商空间理论出发,利用商空间对论域取粒度的方法,导出商空间框架下的大规模的 SVM 数据集的约减模型。约减模型思想如图 1 所示。

在该模型中,约减法的削减幅度随着数据集的缩小而变小,实现约减时最大程度保留原数据集特征。模型思想如下:

第一阶段(粗粒度约减 Coarse-Grained Reduction, CGR):

Step1 设问题集合为 (X, f, T) ,其中 $X=(x, y), x$ 是向量,当 $y=1$ 时称为正类集合,记为 (X_+, f_+, T_+) ,当 $y=-1$ 时称为负类集合,记为 (X_-, f_-, T_-) , $(X, f, T)=(X_+, f_+, T_+) \cup (X_-, f_-, T_-)$;

Step2 分别对 (X_+, f_+, T_+) 和 (X_-, f_-, T_-) 实施某个等价关系 R_1 ,得到两个对应的商空间 $([X_+], [f_+], [T_+])$ 和 $([X_-], [f_-], [T_-])$,分别记为 Q_1 和 Q_2 ;

Step3 对 Q_1 和 Q_2 ,分别再次利用等价关系 $R_2=\{\text{对分}$

类是否有用\},求得 2 个商空间 $([X_+'], [f_+'], [T_+']), ([X_-'], [f_-'], [T_-'])$,分别记为 Q_1', Q_2' ,对分类有用的商空间元素记为 $q_{1\text{有用}}$ 和 $q_{2\text{有用}}$,对分类无用的商空间元素记为 $q_{1\text{无用}}$ 和 $q_{2\text{无用}}$;

Step4 将 $\{q_{1\text{有用}}\}$ 和 $\{q_{2\text{有用}}\}$ 形成两个新的问题集合,记为 X_+' 和 X_-' 。同时,删除 $\{q_{1\text{无用}}\}$ 和 $\{q_{2\text{无用}}\}$;

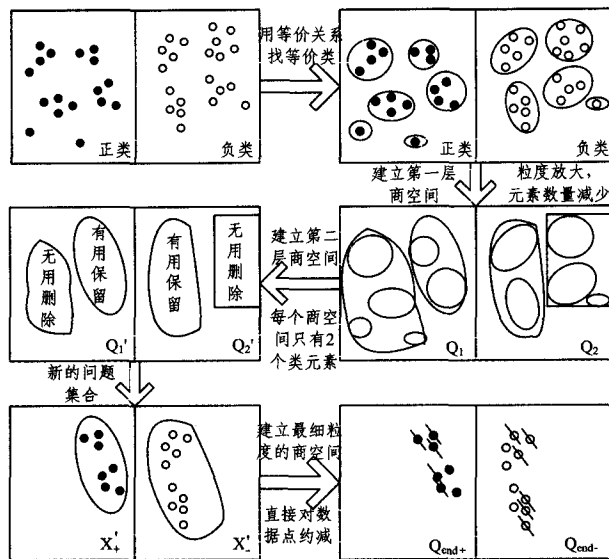


图 1 商空间粒度约减模型思想的示意图

本阶段借助商空间的概念将原问题集合分层,以此找到最合适的删减集。其中,第一步利用某个等价关系 R_1 分别在正类集合 (X_+, f_+, T_+) 和负类集合 (X_-, f_-, T_-) 上生成商空间 Q_1 和 Q_2 ,里面的类元素分别为利用等价关系 R_1 划分出来的等价类。这里将其称为“第一层商空间”。其中等价关系 R_1 需要在算法的实现阶段具体规定。第二步,在第一层商空间上利用等价关系 $R_2=\{\text{对分类是否有用}\}$ 再次建立第二层商空间,即 Q_1', Q_2' 。这里将其称为“第二层商空间”。在第二层商空间里,等价关系 R_2 将第一层商空间的类元素颗粒进一步放大,使得整个商空间里只包含两个类元素:有用类和无用类。最后将第二层商空间中的无用类元素都归入删减集,有用类元素归入删减集补集。粗粒度约减示意图如图 2 所示。

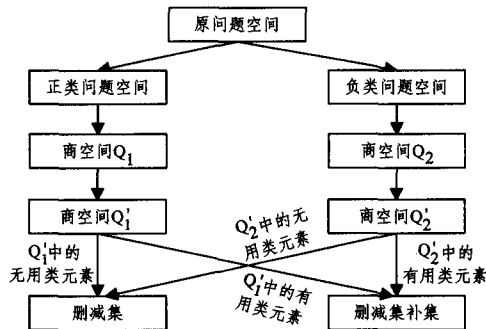


图 2 商空间粗粒度约减模型的示意图

将删减集中的元素去除,对其补集中的元素进行第二阶段的约减。由于第一阶段商空间中的类元素颗粒较大,因此称这一阶段为“粗粒度约减”。

第二阶段(细粒度约减, Fine-Grained Reduction, FGR):

Step1 在新问题集合 X_+' 和 X_-' 上用最细的粒度建立商空间 Q_{end+} 和 Q_{end-} ,在 Q_{end+} 和 Q_{end-} 上讨论约减。

Step2 分别对 Q_{end+} 和 Q_{end-} 中的元素进行筛选。筛选的标准仍然是“对分类是否有用”。保留分类有用的类元素 $\{Q_{end+有用}\}$ 和 $\{Q_{end-有用}\}$, 删除对分类无用的类元素 $\{Q_{end+无用}\}$ 和 $\{Q_{end-无用}\}$ 。

Step3 将 $\{Q_{end+有用}\}$ 和 $\{Q_{end-有用}\}$ 合并后作为约减后的最终集合 X' 。

第二阶段中的新问题集合是第一阶段中删减集的补集, 其中 X_+ 是集合中的正类元素, X_- 是集合中的负类元素。由于数据集的缩小, 这一阶段建立的商空间粒度要加以控制, 这里采用了最小的粒度 R_{end} 对集合进行粒度划分。所谓最小粒度指的是: 操作的粒度不再是等价类, 而是单点数据。这个阶段每个数据点就是一个等价类, 划分的颗粒最小, 因此称为“细粒度约减”。其中提到的筛选标准将在算法的实现阶段具体规定。细粒度约减示意图如图 3 所示。

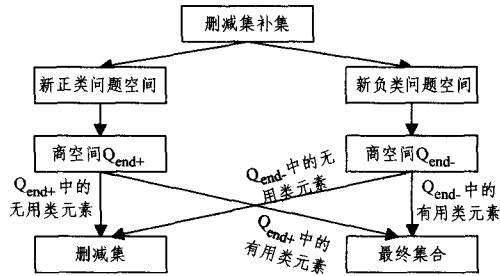


图3 商空间细粒度约减模型的示意图

4 商空间约减法优化大规模 SVM 数据集

针对前文提出的商空间框架下的约减模型思想, 现给出模型的一种实现——双层减样法^[14] (Bistratal Reduction, BR)。该减样法体现了“粒度”的概念, 由强变弱地变幅约减数据集。商空间框架模型与双层减样约减法的转换对照图如图 4 所示。

算法 1 (基于商空间的双层减样算法)

输入: 原始训练样本集 M , 聚簇个数 m , 核函数 $K(x, y)$, 距离参数 D
输出: 约减后的样本集合 M^*

Step1 设当前训练集为 M , 把其中的正类集合记为 T_1 , 负类集合记为 T_2 , $T_2 = M - T_1$;

Step2 把 T_1, T_2 分别用核 k -means 聚类算法聚类得到 m 个簇, 记录下各簇中心;

Step3 利用核距离公式

$$d(x, y) = \|\phi(x) - \phi(y)\| = \sqrt{k(x, x) - 2k(x, y) + k(y, y)} \quad (1)$$

计算 T_1 中各簇到 T_2 中各簇的距离 d_i 和 T_2 中各簇到 T_1 中各簇的距离 d_j , 其中

$$d_i = \sum_{j=1}^m d(x_i, y_j) \quad (2)$$

$$d_j = \sum_{i=1}^m d(y_j, x_i)$$

Step4 挑选在 T_1 中且离 T_2 各簇平均距离最短的前 m 个簇和在 T_2 中且距离 T_1 各簇平均距离最短的前 m 个簇;

Step5 提取 Step4 中挑选出来的样本作为新的训练集 M' ;

Step6 将 M' 代入 SMO 算法求出最优近似解 $\alpha^* = (\alpha_1^*, \dots, \alpha_N^*)^T$;

Step7 根据最优近似解向量 $\alpha^* = (\alpha_1^*, \dots, \alpha_N^*)^T$ 各分量的取值情况, 将大于 0 的分量对应应在 M' 中的训练点挑出, 放入集合 M^* 中。

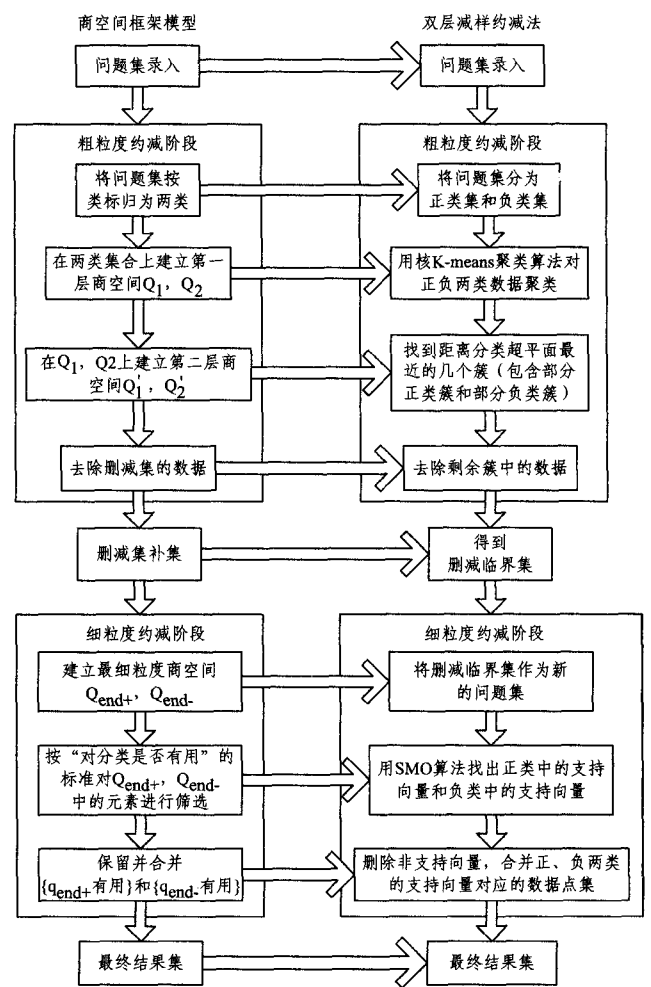


图4 商空间框架模型与双层减样约减法的转换对照图

5 实验

本文实验所涉及的数据集共 3 个, 其中两个对比数据集来自 UCI: EailSpam (ES)、Mushroom (MR), 另一个是利用 PKDD2008 提供的垃圾标签检测数据集处理后生成的 Spammer(SM) 特征数据集, 这是本文的主题数据集。对于每个数据集都将数据标度到 $[0, 1]$ 的范围内, 数据集情况如表 1 所列。

表1 数据集情况

数据集名称	训练集正类数	训练集负类数	训练集总数	训练集属性数
ES	1637	1359	2996	58
MR	3683	2410	6093	22
SM	29217	2461	31677	353063

实验硬件环境: CPU 为 P4, 3.00GHz, 512M 内存。算法实现语言为 C++。本文共设计两个实验, 分类检测环节借助 libSVM 算法包完成, 其中涉及的核函数均选用径向基函数(RBF):

$$K(x, y) = \exp(-\gamma \|x - y\|^2) \quad (3)$$

实验分两个阶段。实验 1, 将没有体现“粒度”概念的 KDC 减样法和基于商空间的双层减样法(QSR)分别作用于这 3 个数据集, 对比减样效果, 以此验证 QSR 相对于普通 KDC 减样法的优势, 实验中的参数采用 libSVM 的默认参数值。实验 2, 用 QSR 优化基于 SVM 的垃圾标签检测模型^[15], 对比优化前、后模型的处理效率。此时 SVM 分类器参数使

用垃圾检测的工程经验值,其优化设置为: $C = 1000$, $\gamma = 0.0001$ 。

实验 1 中的数据集 ES 反映出了普通 KDC 减样法的不足之处。ES 是个很特殊的数据集,其样本点非常凝聚,即使将样本点映射到了核空间,也很难分出多个簇。实验结果表明,QSR 能有效弥补 KDC 减样法。MR 和 SM 的样本点凝聚力没有 ES 强,属于正常样本集。普通数据集 QSR 能在 KDC 减样法的基础上实现进一步无损压缩。实验 1 的结果如表 2 所列。

表 2 两种约减方法效果的对比

数据集 名称	原规模		KDC 减样法		QSR	
	数量	精度(%)	压缩比(%)	精度(%)	压缩比(%)	精度(%)
ES	2996	83.1776	0	83.1776	17.19	83.1776
MR	6093	73.7568	11	84.7858	66.91	84.7858
SM	31677	97.4518	6.4	97.5356	40.32	97.5356

对于凝聚力强的 ES 数据集,由于核空间聚类效果不好,正负类边界簇就是原正负样本集合,无从挑选,因此使用 KDC 减样法,其压缩比为 0,数据集没有改变,因此分类精度没有变化,仍然是 83.1776%。但使用 QSR 约减,数据集得到了进一步压缩。因为 QSR 比 KDC 减样法多了细粒度约减的阶段。当数据集过于凝聚时,簇层次的约减粒度过大,无法进行。此时转用细粒度约减,缩小约减幅度。实验证明,细粒度约减阶段仅保留支持向量对应的样本点,不会损失分类精度。对于普通数据集如 MR 和 SM,用 QSR 可以得到更好的约减效果。因为对于凝聚力一般的数据集,只要将原样本点映射到核空间就能放大数据特征,实现更精准的聚类。这样约减时能准确地去除原数据集的冗余数据,并出现约减后仍有分类精度提高的现象。MR 经 QSR 处理后精度提高 11.029, SM 数据集经优化后精度也提高了 0.0838。不同的数据集情况,精度提高的幅度也不同。MR 经混合法修剪后压缩比比 KDC 减样法提高了 55.91, SM 经混合法修剪后压缩比比 KDC 减样法提高了 33.91。QSR 能从粗细粒度约减两个层面对原样本集做全面修剪。先用粗粒度约减放大数据特征,提高分类精度,再用细粒度约减,挑选保留簇中支持向量的对应点,无损约减。因此,QSR 的实验结果中既能看到压缩比的提高,也能看到精度的提高。总之,QSR 能在 KDC 减样法的基础上得到更好的效果。

实验 2 将原始垃圾标签检测模型^[15]、用 KDC 减样法约减后的检测模型(KDC-SVM)、用基于商空间的双层减样法约减后的检测模型(QSR-SVM)分别带入检测模型完成分类检测任务,并测算所用的训练时间。虽然检测模型用时分为训练用时和预测用时,但预测时间只与支持向量个数有关,与冗余数据多少无关,所以实验 2 中只用训练时间衡量垃圾标签检测模型的优化程度。优化效果对比如图 5 所示,从左边起第一个数据点为没有经过优化的模型训练用时。从图中非常直观地看出此时的模型速度最慢,训练耗时 2223 秒。其次是使用 KDC 减样法优化后的检测模型,训练耗时 1881 秒。效果最好的是使用混合法优化后的检测模型,用时仅有 1178 秒,比未优化过的模型快了 1045 秒。

实验表明,QSR 在 KDC 减样法的基础上进一步实现了无损压缩。在应用方面,将 QSR 用到基于 SVM 的垃圾标签检测模型中,在不损失精度的情况下,提高了训练集的压缩

比,从而明显地缩短了 SVM 的训练时间。这大大优化了原始的检测模型。

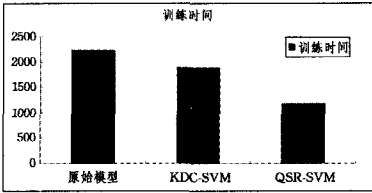


图 5 基于 SVM 的垃圾标签检测模型优化效果对比图

结束语 本文将商空间粒度分析理论用于约减法,把“粒度”的概念引入约减模型,提出商空间框架下的大规模 SVM 数据集约减模型。同时,文中给出了该模型的一种实现——双层减样法。该减样法利用“粒度”概念把聚类约减和单点约减有机结合,形成了效果更好的 SVM 约减法,为解决大规模支持向量机的学习问题提供了新的解决途径。

参 考 文 献

[1] Rychetsky M,Ortmann S,Ullmann M,et al. Accelerated Training of Support Vector Machines[C]//International Joint Conference on Neural Networks,1999. IEEE,1999:998-1003

[2] Hsu C-W, Lin C-J. A Comparison of Methods for Multiclass Support Vector Machines[J]. Neural Networks, 2002, 13(2): 415-425

[3] Yu H,Yang Jiong,Han Jia-wei. Classifying Large Data Sets Using SVMs with Hierarchical Clusters[C]//Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2003. New York:ACM,2003:306-315

[4] 奉国和,朱思铭. 基于聚类的大样本支持向量机研究[J]. 计算机科学,2006,33(4):145-147

[5] Xu Yan-zi, Qin Hua. A New Optimazation Method of Large-scale SVMs Based on Kernel Distance clustering[C]//Computational Intelligence and Software Engineering, 2009. Wuhan: IEEE,2009

[6] Cevikalp H. New clustering algorithms for the support vector machine based hierarchical classification[J]. Pattern Recognition Letters,2010,31(11):1285-1291

[7] He Qiang, Xie Zong-xia, Hu Qing-hua. Neighborhood based sample and feature selection for SVM classification learning[J]. Neurocomputing, 2011, 74(10):1585-1594

[8] 胡正平,高文. 基于改进加权压缩近邻与最近边界规则 SVM 训练样本约减选择算法[J]. 燕山大学学报,2010,34(5):421-425

[9] 张钊,张铃. 粒计算未来发展方向探讨[J]. 重庆邮电大学学报:自然科学版,2010,22(5):538-540

[10] 郭庆文,王国胤,张清华. 多粒度的图像检索方法研究[J]. 计算机科学,2013,40(1):298-301

[11] 程伟,石扬,张燕平. 粒度计算的三种主要方法[J]. 计算机技术与发展,2007,17(3):91-94

[12] 李道国,苗夺谦,张红云. 粒度计算的理论、模型与方法[J]. 复旦学报:自然科学版,2004,34(5):837-841

[13] 张玲,张钊. 模糊商空间理论[J]. 软件学报,2003,14(4):770-776

[14] 覃希,苏一丹. 用双层减样法优化大规模 SVM 垃圾标签检测模型[J]. 计算机应用研究,2011,28(6):2095-2098

[15] 覃希,夏宁霞,苏一丹. 基于支持向量机的垃圾标签检测模型[J]. 计算机应用研究,2010,27(10):3893-3895