文章编号:1001 - 9081(2009)10 - 2736 - 05

一种新的支持向量机大规模训练样本集缩减策略

朱 方[1],顾军华[2],杨欣伟[2],杨瑞霞[1]
(1. 河北工业大学 信息工程学院,天津 300401; 2. 河北工业大学 计算机科学与软件学院,天津 300401)
(sky050607@sina.com)

摘 要:支持向量机(SVM)在许多实际应用中由于训练样本集规模较大且具有类内混杂孤立点数据,引发了学习速度慢、存储需求量大、泛化能力降低等问题,成为直接使用该技术的瓶颈。针对这些问题,通过在点集理论的基础上分析训练样本集的结构,提出了一种新的支持向量机大规模训练样本集缩减策略。该策略运用模糊聚类方法快速的提取出潜在支持向量并去除类内非边界孤立点,在减小训练样本集规模的同时,能够有效地避免孤立点数据所造成的过学习现象,提高了 SVM 的泛化性能,在保证不降低分类精度的前提下提高训练速度。

关键词:支持向量机;点集;模糊 C-均值;潜在支持向量;孤立点
中图分类号: TP181 **文献标志码**:A

New reduction strategy of large-scale training sample set for SVM

ZHU Fang[1], GU Jun-hua[2], YANG Xin-wei[2], YANG Rui-xia[1]
(1. *School of Information Engineering, Hebei University of Technology, Tianjin* 300401, *China*;
2. *School of Computer Science and Software, Hebei University of Technology, Tianjin* 300401, *China*)

Abstract: It has become a bottleneck to use Support Vector Machine (SVM) due to such problems as slow learning speed, large buffer memory requirement, low generalization performance and so on, which are caused by large-scale training sample set and outlier data immixed in the other class. Concerning these problems, this paper proposed a new reduction strategy for large-scale training sample set according to the analysis on the structure of the training sample set based on the point set theory. This new strategy gets the potential support vectors and removes the non-boundary outlier data immixed in the other class by using fuzzy clustering. That can greatly reduce the scale of the training sample set and improve the generalization performance by effectively avoiding over-learning caused by outlier data, and finally speed up learning rate without reducing the classification accuracy.

Key words: Support Vector Machine (SVM); point set; Fuzzy C-Means (FCM); potential support vector; outlier

0 引言

支持向量机(Support Vector Machine,SVM)是 Vapnik 等人根据统计学习理论提出的一种机器学习方法。由于它能够有效地避免局部极小值,且具有良好的推广性能和较好的分类精确性,所以,近年来在模式识别、回归分析和特征提取等方面得到了越来越广泛的应用,已经成为国际上人工智能领域和机器学习领域新的研究热点。然而在许多实际学习训练中由于学习样本集很大,造成学习速度慢,存储需求量大,成为直接使用 SVM 技术的障碍;并且,对于在相对类样本中混杂有孤立点数据的样本集进行训练时,往往无助于提高分类器的性能,反而会大大增加训练器的计算负担,同时它们的存在还可能造成过学习,从而增大了分类判别函数的 VC 维,使置信区间变大,最终影响 SVM 的泛化能力,为此出现了许多改进的支持向量机算法[1-8]。

文献[6,8]中提出的缩减策略是基于类中心思想提出的,在得到原空间中正负样本聚类中心的基础上,通过判定样本与聚类中心的规定半径之间的关系来实现对训练样本集的缩减;但该种分析策略只适合与正负样本集都为凸集的情况,否则无效。文献[4]的作者通过 C-均值聚类方法对训练样本集进行分组,如果一组中所有样本都来自于同一类,则用聚类中心代替,否则,则保留该组中所有的样本;该方法虽然能够对非凸集训练样本集进行有效缩减,但是当聚类数低于样本数的 1/20 时,缩减效果极为不明显,对于大规模样本集,随聚类数目的增加,将会以计算时间的增加为代价换取训练样本集数目的减少,使算法不具备实际意义。文献[5]的作者提出的 NN-SVM 算法,根据每个样本与其最近邻类标的异同决定其取舍,在降低样本集规模的同时能够减少孤立点数据对 SVM 泛化性能的影响;但在寻找每个样本点的最近邻点时将耗费大量的时间,对于大规模样本集来说,该算法的效率极低,也失去了实际意义。文献[3]的作者提出另一种缩减策略 PSCC,根据训练样本在高维空间中线性可分的几何特性,通过计算高维空间中每个样本与正负聚类中心连线的夹角来实现对样本集的缩减,是最近提出的具有实际意义的一种算法;但该方法需要大量的核计算,效率不高。

通过分析上述现有改进算法的基本思想及其优缺点,本文针对含有类内混杂孤立点数据的大规模训练样本集,提出了一种基于点集理论的支持向量机大规模训练样本集缩减策略(Reduction Strategy for SVM Large-Scale Training Sample Set, SVM-LSTSRS),在大量减小训练样本集规模的基础上,能够有

收稿日期:2009 - 04 - 16。 **基金项目**:天津市自然科学基金资助项目(07JCZDJC10800)。
作者简介:朱方(1981 -),男,河北秦皇岛人,博士研究生,主要研究方向:模式识别、智能系统; 顾军华(1966 -),男,河北石家庄人,教授,博士生导师,主要研究方向:数据挖掘、智能信息处理; 杨欣伟(1984 -),女,河北唐山人,硕士研究生,主要研究方向:模式识别; 杨瑞霞(1957 -),男,河北清河人,教授,博士生导师,主要研究方向:微电子新器件及新材料。

效地减少混杂在相对类中的孤立点数据对分类判别函数的影响,从而既提高训练速度,又不影响 SVM 的分类性能。

1　支持向量机(SVM)

SVM 算法[9]是从线性可分情况下的最优分类超平面提出的,并把此平面作为最终分类决策面。该算法将寻找最优分类超平面转换成二次凸规划问题,保证了能够得到全局最优解。同时它还根据 Mercer 条件引入核函数,将低维空间中线性不可分的样本映射到高维空间中的线性可分样本,并巧妙地解决了映射带来的"维数灾难"问题,实现了对非线性可分样本的准确分类。

1.1　最优分类超平面

把问题限定在线性可分情况下,最优分类超平面就是要求分类线不但能将两类无错误地分开,而且要使两类的分类间隔最大。

设线性可分样本集为(x_i,y_i),$i=1,2,\cdots,n$,$x\in\mathbf{R}^d$,$y\in\{+1,-1\}$是类别号,d维空间中的分类超平面方程为:

$$\boldsymbol{\omega}\cdot\boldsymbol{x}+b=0 \tag{1}$$

进行归一化后,得分类间隔为:

$$\frac{|1-b+1+b|}{\|\boldsymbol{\omega}\|}=\frac{2}{\|\boldsymbol{\omega}\|} \tag{2}$$

依据最优分类超平面的要求,问题转化为满足式(3)条件下:

$$y_i[(\boldsymbol{\omega}\cdot\boldsymbol{x})+b]-1\geqslant 0;\ i=1,2,\cdots,n \tag{3}$$

求式(4)的最小值:

$$\phi(\boldsymbol{\omega})=\frac{1}{2}\|\boldsymbol{\omega}\|^2=\frac{1}{2}(\boldsymbol{\omega}\cdot\boldsymbol{\omega}) \tag{4}$$

转化为 Lagrange 对偶问题,即在下述条件下:

$$\sum_{i=1}^{n}y_i\alpha_i=0;\alpha_i\geqslant 0,i=1,2,\cdots,n \tag{5}$$

求如式(6)所示的函数极大值:

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \tag{6}$$

可知,对偶问题完全是根据训练数据来表达的,函数Q的最大化仅依赖于输入样本点积的集合。

根据 Karush-Kuhn-Tucker 条件,优化问题的解满足:

$$\alpha_i(y_i(\boldsymbol{\omega}\cdot x_i+b)-1)=0;i=1=2,\cdots,n \tag{7}$$

因此,对多数样本α_i^*将为零,取值不为零的α_i^*对应于上式等号成立的样本即支持向量,它们通常只是全体样本中很少一部分。

对于确定的空间,有定理 1 成立,这是进行样本集缩减的理论依据。

定理 1　支持向量机的训练结果与非支持向量无关。

1.2　核函数

低维空间中的样本集往往难以划分,可以映射到高维空间$\varphi(x)$,使之线性可分。根据 Mercer 条件,若函数$K(x,x')$满足 Mercer 条件,可以用核函数$K(x,x')$代替高维空间中的点积,即$K(x,x')=\varphi(x)\cdot\varphi(x')$,这样,只要适当选取核函数,就可以得到对应的高维空间中的目标函数:

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{8}$$

通过 SVM 的基本原理可以看到,由于最终对分类起贡献的样本仅仅是支持向量,但其在整个样本集中占的比例却很小,因此支持向量机对小规模训练集非常有效,但实际中训练集规模常常又比较大,在解决二次规划的问题中,训练的迭代过程需要多次使用 Hessian 矩阵,以至耗费大量的时间,且占用相当大的内存容量,所以需要对样本集进行缩减,来提高支持向量机的学习效率。

2　支持向量机大规模训练样本集缩减策略

针对 SVM 对大规模训练样本集进行训练时所存在的固有缺陷,本文从点集的角度出发,通过分析单个样本点与相对类样本集的关系,对潜在支持向量(两类边界附近的包含支持向量在内的一个较小点集)进行了准确定位,并通过运行效率较高的模糊 C 均值聚类算法所得到的模糊隶属度矩阵对样本点与相对类样本集的关系进行判定,最终有效地去除大部分非支持向量和部分混杂在另一类中的孤立点数据,得到潜在支持向量。

2.1　SVM-LSTSRS 原理

SVM-LSTSRS 是在点集理论基础上提出的,首先给出点集的定义。

定义 1　集合中元素间有某种关系、集合内有某种结构的集合,叫做点集[10]。

设E是n维空间$\mathbf{R}^n$中的一个有限点集,P_0是$\mathbf{R}^n$中的一个定点,P_0与E的关系有三种互斥的情形:1)P_0附近根本没有E的点;2)P_0附近全是E的点;3)P_0附近既有E的点,又有不属于E的点。

针对上述情形给出如下定义。

定义 2　如果存在P_0的某一邻域$U(P_0)$,使$U(P_0)\subset E$,则称P_0为E的内点。

定义 3　如果P_0是$\overline{E}$的内点(这里补集是对全空间$\mathbf{R}^n$来说的),则称P_0为E的外点。

定义 4　如果P_0既非E的内点又非E的外点,也就是:P_0的任一邻域内既有属于E的点,也有不属于E的点,则称P_0为E的边界点。

这里,把两类d维训练样本集$X_1=(x_1,\cdots,x_{n1})$和$X_2=(x_1,\cdots,x_{n2})$,类别分别为+1和-1,看作是$\mathbf{E}^d$空间中的两个点集E_1和E_2。根据点集的定义,我们可知E_1中的任意一点p与点集E_2的关系只有三种,即点p为E_2的外点、内点或边界点。

如果点p为E_2的外点,即点p周围无E_2中的点,那么点p不是处于两类边界上的点,为非支持向量;如果点p为E_2的内点(混杂在另一类中的孤立点数据),即点p周围全是E_2中的点,则点p也不是处于两类边界上的点,为非支持向量;如果E_1中的点p为E_2的边界点,即点p周围既有非E_2中的点又有E_2中的点,那么点p是处于两类边界上的点,为潜在支持向量。同样E_2中的点与点集E_1的关系也有上述3种。所以,我们只需找到两类的边界点,就能得到潜在支持向量。

2.2　基于 FCM 的样本点类型判定

在单个样本点类型判定过程中,按照常规方法,需要计算每两个点之间的距离,来得到该点的近邻点,但这需要$O(n^2)$阶的距离计算,对于大规模样本,需要耗费大量的时间。这里,我们通过一种基于划分的模糊聚类算法将训练样本划分为c类,并用每类的聚类中心作为每个子类所有点的代表,完成对p点类型的判定,其中,要求用来作为判定依据的聚类中心必须有明确的所属类别。

由于模糊 C 均值(Fuzzy C-Means,FCM)聚类具有算法思想简单,易实现,而且收敛速度快,有着较高的执行效率等优点,所以选择其作为得到样本点类型判别依据的方法。

FCM 把 n 个向量 $\boldsymbol{x}_i(i=1,2,\cdots,n)$ 分为 c 个模糊类,并求每类的聚类中心 $c_i(i=1,2,\cdots,n)$,使得非相似性指标的价值函数达到最小,最终得到模糊隶属度矩阵。算法中应用模糊划分,使得每个给定数据点用值在[0,1]内的隶属度来确定其属于各个类的程度,一个数据集的隶属度和总等于1:

$$\sum_{i=1}^{c} u_{ij} = 1; \forall j = 1,\cdots,n \tag{9}$$

价值:

$$J(U,c_1,\cdots,c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^m d_{ij}^2 \tag{10}$$

其中:$d_{ij} = \| x_{ij} - c_i \|$ 为第 j 个数据点与第 i 个聚类中心间的距离,$m \in [1,\infty)$ 是一个加权指数。

针对所得到隶属度矩阵,取 E_1 中的样本点 p 做如下判定规则(c_1 为聚类中心类别为 +1 的聚类中心个数,c_2 为聚类中心类别为 -1 的聚类中心个数,u_{ip} 为点 p 到第 i 个聚类中心的隶属程度):

1)若 $\sum_{i=1}^{c1} u_{ip} - \sum_{i=1}^{c2} u_{ip} \geqslant \lambda$,其中 $0 < \lambda \leqslant 1$,则 p 为 E_2 的外点(非支持向量);

2)若 $\sum_{i=1}^{c2} u_{ip} - \sum_{i=1}^{c1} u_{ip} \geqslant \lambda$,其中 $0 < \lambda \leqslant 1$,则 p 为 E_2 的内点(混杂在另一类中的非边界孤立点);

3)若 $\left| \sum_{i=1}^{c2} u_{ip} - \sum_{i=1}^{c1} u_{ip} \right| < \lambda$,其中 $0 < \lambda \leqslant 1$,则 p 为 E_2 的边界点(潜在支持向量)。

2.3 SVM-LSTSRS 实现步骤

根据上面提出的算法基本思想,SVM-LSTSRS 实现步骤如下:

1)给定训练样本集 E_1($h1$ 组)和 E_2($h2$ 组),并给定每类的初始聚类中心数目 $m\left(m = 2\dfrac{(h1+h2)}{1000}\right)$;

2)针对总样本集 $E_1 + E_2(x_1,x_2,\cdots,x_n)$,根据式(11),计算隶属度矩阵;

$$u_{ij} = \begin{cases} \dfrac{1}{\sum_{k=1}^{c} \left(\dfrac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}}, & d_{ij} \neq 0 \\ 1, & d_{ij} = 0, j = k \\ 0, & d_{ij} = 0, j \neq k \end{cases} \tag{11}$$

其中:n 为样本集总数,$c = 2m$,d_{ij} 为样本点 $\boldsymbol{x}_i$ 与样本点 x_j 之间的欧式距离。

3)计算价值函数 J,如果它小于某个确定的阈值或它相对上次价值函数值的改变量小于某个阈值,则迭代停止,得到最终的聚类中心 C、模糊隶属度矩阵 $\boldsymbol{U}$ 和距离矩阵 $\boldsymbol{D}$,转5);否则,转4);

4)根据式 $c_i = \dfrac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$ 重新计算聚类中心 $c_i, i = 1,2,\cdots,c$,返回2)。

5)根据 k 近邻判别法则$\left(\text{其中 } k = \dfrac{(h1+h2)}{20}\right)$,判定各个聚类中心所属的类别 +1 或 -1,去掉类别无法判定的聚类中心对模糊隶属度矩阵的影响,即将隶属度矩阵中该聚类中心对应行的所有元素置0;

6)对模糊隶属度矩阵进行修正,保证类别为 +1 和 -1 的聚类中心数均衡且各个点对同一聚类中心的隶属度总和仍为1。其中修正原则为:分别统计类别标识为 +1 和 -1 的聚类中心总数 sum_1 和 sum_2,如果 $sum_1 \neq sum_2$,则按式(12)将隶属度矩阵 $\boldsymbol{U}$ 补充 $|sum_1 - sum_2|$ 行:

$$u_{ij} = \min\{u_{kj} \mid y_{ck} = \operatorname{sgn}(sum_1 - sum_2)\} \tag{12}$$

$i = c+1,\cdots,c+|sum_1 - sum_2|$; $k = 1,\cdots c$,y_{ck} 为经 FCM 得到的第 k 个聚类中心的类别标识。最后,规范化隶属度矩阵:

$$u_{ij} = \frac{1}{c+|sum_1 - sum_2|} u_{ij} \Big/ \sum_{k=1} u_{kj} \tag{13}$$

7)根据2.2节中的规则去除非支持向量和非边界混杂孤立点,得到潜在支持向量。

2.4 SVM-LSTSRS 性能分析

在不进行预选时,对原样本集(样本规模为 n)进行 SVM 训练时,迭代过程需要多次使用 Hessian 矩阵,对核函数需要进行 $O(n^2)$ 阶的计算。在核函数复杂度比较高时,计算量很大,并且样本集规模比较大时,二次凸规划的寻优过程中需要的迭代过程很多,最终将耗费大量的时间,导致算法的效率很低。

在 SVM-LSTSRS 中,通过 FCM 聚类过程来缩减大量的非支持向量,大大降低了进行训练的样本规模。FCM 聚类过程需要进行 $O(iter \times c \times n)$ 阶的欧氏距离计算,一般我们规定 $c \ll n$,且 $iter \ll 100$(其中,c 为聚类数,$iter$ 为最大迭代次数),则经过预选后的潜在支持向量数量为 n_1,需要进行 $O(n_1^2)$ 阶的核函数计算,正常情况下,支持向量只占训练样本集边界数据中的极少数,所以可以认为 $n_1 < n/2$,故如果只考虑上述计算消耗的时间,可得 $O(n^2)$ 小于 $O(iter \times c \times n) + O(n_1^2)$,该缩减策略是具有实际意义的。

其次,在预选策略中还能够有效去除部分混杂在另一类样本集中的非边界孤立点,从而避免了不必要的过学习现象,得到适中的置信区间,最终提高了 SVM 算法的泛化性能。

3 实验结果及分析

为了评价本文所提出的 SVM-LSTSRS 的有效性能,将实验分为三种,一是采用二维可视化数据,对算法的缩减结果进行直观的评定;二是采用 Libsvm 网站中提供的数据集进行实验,来证明改进算法的普遍应用性;三是通过与目前应用较多的 Shrinking 缩减方法和最近提出的较为有效且具有实际意义的 PSCC 算法进行对比,来说明该算法的提出具有实际意义。

3.1 二维可视数据

分别对二维空间中在另一类中有混杂孤立点的大规模训练样本集进行 LSTSRS 实验,其中,图1和2为线性可分情况,图3和4为线性不可分情况,实心点和 + 点分别代表两类不同样本,空心圆圈标记的为经 LSTSRS 得到的潜在支持向量。

图3中实心点满足方程1:$(x-2)^2 + (y-2)^2 \leqslant 1$,图4

中实心点满足方程2:$y-0.2\geqslant(x-0.5)^2$。

从图1~3可以看出,预选策略能够有效去除大部分非两类边界的样本点和部分混杂在另一类中的非边界孤立点数据,大大减少下一步进行SVM学习训练的样本集规模,且避免了不必要的过学习现象,提高了SVM的泛化性能。

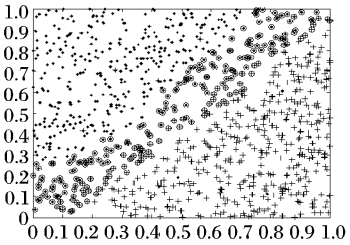

图1　1000组二维均匀分布的样本点

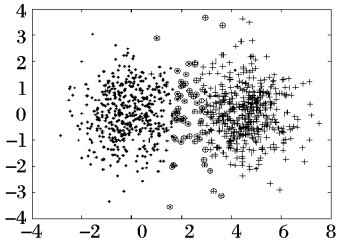

图2　1000组二维正态分布的样本点

3.2　Libsvm提供的分类测试数据

采用Libsvm网站上提供的UCI、CWH03a和DP01a数据库进行仿真实验,硬件环境为CPU-AMD5200,1 GB内存,开发环境为Libsvm软件包和C++6.0编译器。为了避免环境和起始聚类中心随机选取造成的差异,实验结果均是5次取均值。实验选择RBF核函数,不采用交叉验证,结果见表1,其中:N_{tr}代表训练样本集规模,N_t代表测试样本集规模,T_{tr1}代表不经LSTSRS所需要的训练时间,N_{svm1}代表不经LSTSRS进行训练所得到的支持向量个数,P_1代表对测试样本集进行预测的准确率,λ为缩减策略中的规定值,T_{sel}为对样本集进行缩减所消耗的时间,N_q为得到的潜在支持向量个数,P_r为所得到的潜在支持向量中含有支持向量的准确率,T_{tr2}为对潜在支持向量的训练时间,N_{svm2}为对潜在支持向量进行训练所得到的支持向量个数,P_2代表用缩减后样本集经训练所得到的支持向量对测试样本集进行预测的准确率,r代表LSTSRS+SVM的时间缩减率($r=(T_{tr1}-T_{sel}-T_{tr2})/T_{tr1}$)。

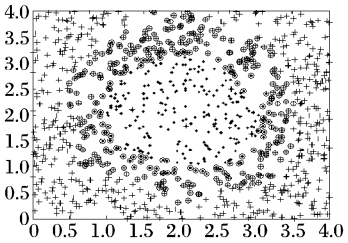

图3　1000组线性不可分样本点(方程1)

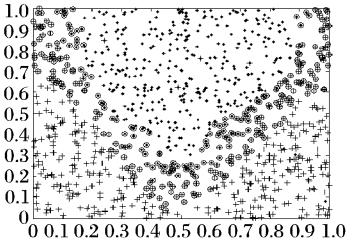

图4　1000组线性不可分样本点(方程2)

表1　LSTSRS前后的对照表

Source	data	dim	N_{tr}	N_t	*SVM*			LSTSRS+SVM						
					T_{tr1}/s	N_{svm1}	P_1/%	λ	T_{sel}/s	N_q	T_{tr2}/s	P_r/%	N_{svm2}	P_2/%	r/%
CWH03a	svmguid1	4	5089	2000	0.5620	591	92.2500	0.4	0.0280	1519	0.1560	89.51	512	92.8500	67.26
								0.6	0.0280	2805	0.2650	97.97	554	92.3000	47.86
								0.8	0.0280	4860	0.5050	99.66	588	92.2500	5.10
CWH03a	svmguid3	21	837	447	0.4530	371	84.3400	0.4	0.0203	565	0.1400	80.62	309	83.9843	64.61
								0.6	0.0212	709	0.1710	90.03	334	84.6689	57.57
								0.8	0.0270	801	0.3190	98.11	361	85.0112	23.62
DP01a	ijcnn1	22	7853	880	1.3900	153	99.2727	0.4	0.1560	3756	0.1880	84.05	128	99.3727	75.25
								0.6	0.2035	4312	0.2650	90.36	138	99.7045	66.29
								0.8	0.1950	6278	0.9690	96.49	147	99.2103	16.26

通过上面的实验结果,可知通过对预选得到的潜在支持向量进行训练来预测测试样本集时,识别准确率并不会降低,有的反而会有一定程度的提高;其次,经LSTSRS,可以大大缩小训练样本集的规模,提高SVM的训练效率,如对ijcnn1数据集进行训练时,在$\lambda=0.6$时,可以将整体训练时间缩短66.29%,且此时的分类精度并没有降低。

但随着缩减策略中的规定值λ的增大,潜在支持向量的规模会不断增大,直至接近原训练样本集,可见,在$\lambda\geqslant0.8$时,虽然潜在支持向量中含有原支持向量的准确率会有所提高,但算法效率的优越性不能得到很好的体现,且对测试样本的预测准确率没有明显的提高。由上述分析可知λ取值为0.6左右时,LSTSRS的效果最为明显,可以作为经验参考值。

3.3　LSTSRS与其他算法比较

虽然目前出现了很多对样本缩减的方法,但在实际中应用最多的,还是Joachims提出的Shrinkin方法,目前流行的SVMligth、LIBSVM、SVMTorch等软件包都使用了shrinking技术缩减样本,是一种较为有效的缩减技术。其次,PSCC算法是最近新提出的一种缩减方法,与最近提出的其他方法相比,更具有代表性和坚实的理论基础。因此本文通过和这两种方法的实验结果进行对比,来说明LSTSRS的有效性和实际应用价值。实验环境同上。采用a5a和a6a数据集进行实验,LSTSRS方法中的参数$\lambda=0.6$,PSCC方法中的参数$\gamma=0.10$,$\varepsilon=-0.1$。表2中的dim表示样本维数,N_{tr}表示训练样本规模,N_t表示测试样本规模,T_{tr}表示训练总时间(包括样

本缩减时间),N_{sv} 表示支持向量数,P_r 表示预测正确率。

通过表2中的实验数据可以看出,LSTSRS较Shrinking技术和PSCC算法有了更高的效率,且通过有效去除类内混杂的非边界孤立点数据,提高了SVM的泛化性能,使分类正确率有了一定程度的提高,故该缩减策略是有意义的,且符合实际应用的要求。

表2 LSTSRS方法与PSCC和Shrinking方法的比较

Data	dim	N_{tr}	N_t	SVM			Shrinking+SVM			PSCC+SVM			LSTSRS+SVM		
				T_{tr}/s	N_{sv}	P_r/%	T_{tr}/s	N_{sv}	P_r/%	T_{tr}/s	N_{sv}	P_r/%	T_{tr}/s	N_{sv}	P_r/%
a5a	123	5318	7000	76.875	2330	74.0857	39.859	2338	74.0857	35.7521	2311	73.6849	32.3516	2228	75.8320
a6a	123	11220	9945	281.249	4004	79.2459	139.141	4004	79.2459	106.1847	3911	77.7329	95.3830	3814	79.2547

4 结语

本文提出的SVM-LSTSRS从分析单个样本点与相对样本集合的关系出发,给出了潜在支持向量的概念,并以得到潜在支持向量为目的进行样本集的缩减。在LSTSRS中为了避免大量的计算负担,引入了模糊C-均值聚类,并在得到隶属度矩阵的基础上对样本点与另一类样本集的相对关系给出了合理的判定规则。通过实验证明该缩减策略在保证分类精度的同时能够大大缩小训练样本集的规模,并降低了孤立点对SVM泛化性能的影响,具有较高的执行效率,是有效且可行的。

参考文献:

[1] AGARWAL D K. Shrinkage estimator generalizations of proximal support vector machines[C]// Proceedings of the 8th ACM SIGKDD International Conference of Knowledge Discovery and Data Mining. New York: ACM Press, 2002:173-182.

[2] DANIAEL B, CAO D. Training support vector machines using adaptive clustering[C/OL].[2009-02-01]. http://www.siam.org/proceedings/datamining/2004/dm04_012boleyd.pdf.

[3] 罗瑜.支持向量机在机器学习中的应用研究[D].成都:西南交通大学,2007.

[4] 肖小玲,李腊元,张翔.提高支持向量机训练速度的CM-SVM方法[J].计算机工程与设计,2006,27(22):4183-4184.

[5] 李红莲,王春花,袁保宗.一种改进的支持向量机NN-SVM[J].计算机学报,2003,26(8):1015-1020.

[6] 曾志强.支持向量机分类机的训练与简化算法研究[D].杭州:浙江大学,2007.

[7] 谭冠群,丁华福.支持向量机方法在文本分类中的改进[J].信息技术,2008,2(1):83-88.

[8] 曹淑娟,刘小茂,张钧,等.基于类中心思想的去边缘模糊支持向量机[J].计算机工程与应用,2006,42(22):146-149.

[9] CRISTIANINI N, SHAWE-TAYLOR J. An introduction to support vector machines and other kernel-based learning methods[M].李国正,王猛,曾华军,译.北京:电子工业出版社,2006.

[10] 程其襄,张奠宙,魏国强,等.实变函数与泛函分析基础[M].2版.北京:高等教育出版社,2003:31-37.

(上接第2728页)

表1 算法改进前后PR值及排序

网页	改进前PR值	改进前排名	改进后PR值	改进后排名
F_1	0.4535	1	0.6025	1
F_2	0.4309	2	0.4309	8
F_3	0.3771	3	0.5474	2
…	…	…	…	…
F_7	0.3426	7	0.4837	4
F_8	0.3314	8	0.3746	11
…	…	…	…	…
F_{12}	0.2987	12	0.2987	12
F_{13}	0.2814	13	0.4634	7
F_{14}	0.2792	14	0.2792	14
…	…	…	…	…

由于实验原因,部分相关度不高的网页未被模拟点击。而F_2是关于风姿物语的百度百科,虽然与主题相关但与所求无关,所以被排除掉,因此权值不变。F_1,F_7和F_{13}由于是专门的小说网站上的网页,被用户识别并点击的次数很大,F_{13}由于本身位于第二页,其加权率本身就高,所以在较少的点击率下排名上升到第7位,在排序结果的第一页显示。由表1可以看出,改进后的算法提升了用户真正感兴趣的网页的名次,从而使用户可以更快地找到自己所需的结果。

4 结语

本文从网页相对于关键字的点击率出发,通过蚁群算法的信息熵的概念将用户的群体选择加入到网页权值计算中去,提高了相关网页的查准率。但其本身仍存在一定的问题。如网页加权的权值是否可以继续传递下去,这样会造成一部分相关网页权值的整体上升等,这些问题还需要进一步的研究。

参考文献:

[1] BRINKMEIER M. PageRank revisited[J]. ACM Transactions on Internet Technology, 2006,6(3): 282-301.

[2] RICARDO B-Y, BERTHIER R-N. Modern information retrieval[M]. 王知津,贾福新,郑红军,等译.北京:机械工业出版社,2004.

[3] 黄德才,戚华春,钱能.基于主题相似度模型的TS-PageRank算法[J].小型微型计算机系统,2007,28(3):510-513.

[4] 宋聚平,王永成,尹中航,等.对网页PageRank算法的改进[J].上海交通大学学报,2003,37(3):397-400.

[5] 戚华春,黄德才,郑月锋,等.具有时间反馈的PageRank改进算法[J].浙江工业大学学报,2005,33(3):272-275.

[6] HAVELIWALA T H. Topic-sensitive PageRank[C]// Proceedings of the Eleventh International World Wide Web Conference. Honolulu: ACM Press, 2002:517-526.

[7] RICHARDSON M, DOMINGOS P. The intelligent surfer: Probabilistic combination of link and content information in PageRank[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2002:1441-1448.

[8] ZHE CAO, TAO QIN, LIU TIE-YAN, *et al*. Learning to rank: From pairwise approach to listwise approach[C]// Proceedings of the 24th International Conference on Machine Learning. New York: ACM Press, 2007:129-136.

[9] 王建勇,单松巍,雷鸣,等.海量Web搜索引擎系统中用户行为的分布特征及其启示[J].中国科学:E辑,2001,31(4):372-384.