基于超球支持向量机的类增量学习算法研究*)

秦玉平[1,2] 李祥纳[2] 王秀坤[1] 王春立[1]

(大连理工大学电子与信息工程学院 大连 116024)[1] (渤海大学信息科学与工程学院 锦州 121000)[2]

摘 要 提出了一种超球支持向量机类增量学习算法。对每一类样本,利用超球支持向量机在特征空间中求得包围该类尽可能多样本的最小超球,使各类样本之间通过超球隔开。类增量学习过程中,只对新增类样本进行训练,使得该算法在很小的样本集、很小的空间代价下实现了类增量学习,大大降低了训练时间,同时保留了历史训练结果。分类过程中,通过计算待分类样本到各超球球心的距离判定其所属类别,分类简单快捷。实验结果证明,该算法不仅具有较高的训练速度,而且具有较高的分类速度和分类精度。

关键词 支持向量机,类增量学习,超球

Study on Class Incremental Learning Algorithm Based on Hyper-sphere Support Vector Machines

QIN Yu-ping[1,2] LI Xiang-na[2] WANG Xiu-kun[1] WANG Chun-li[1]

(School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China)[1]

(College of Information Science and Technology, Bohai University, Jinzhou 121000, China)[2]

Abstract A new class incremental learning algorithm based on hyper-sphere support vector machines is proposed in this paper. For every class, hyper-sphere support vector machine is used to get the smallest hyper-sphere in feature space that contains most samples of a class, which can divide the class samples from others. In the process of class incremental learning, only are the samples that belong to the new incremental class trained. Therefore, the class incremental learning can be realized in a small set of samples and a small memory space. The training time is reduced largely; besides, the history results of the classes that have nothing to do with the incremental class are saved at the same time. For the sample to be classified, the distances from it to the centre of every hyper-sphere are used to confirm the class that the sample belongs to. The classification method is simple and the speed is fast. The experiment results show that the algorithm has a higher performance on training speed, classification speed, and classification precision.

Keywords Support vector machine, Class incremental learning, Hyper-sphere

1 引言

支持向量机 SVM(Support Vector Machines)[1]是一种基于统计学习理论的机器学习方法,具有很好的泛化能力。支持向量机的泛化性能并不依赖全部训练数据,而是该全部数据的一个子集,即支持向量,并且支持向量的数目同整个训练数据集数目相比是很小的,因此支持向量机对增量学习是一种强有力的工具。

支持向量机增量学习的主要研究成果有 Batch SVM 增量学习算法[2,3]、训练精确解算法[4]、分块增量算法[5]、快速增量学习算法[6]、α-ISVM 算法[7]、HS-SVM[8]和多分类增量学习算法[9]等,这些方法都不涉及类别的增加。文献[10]提出了一种类增量学习算法,该算法将新增类样本作为正类,原有类样本作为负类,训练一个二值分类器,增加到原分类系统中,并将其作为二叉树的根结点。该算法节省了训练时间,取得了一定的效果。但随着新类别的不断增加,新增类别的训练样本数将会远远少于原有类训练样本总和,致使两类训练样本数量严重不均衡,导致分类精度降低。此外,增量学习过程中,所有样本都参加训练,必将导致计算复杂度过高,当数据增加到一定程度时,训练无法实现。

本文提出了一种基于超球支持向量机的类增量学习算法(HS-CIL)。将样本空间的每一类样本用超球定界,使得特征空间中不同类的样本之间通过超球分开,达到分类效果。类增量学习过程中,只对新增类样本进行训练,使得该算法在很小的样本集、很小的空间代价下实现了类增量学习,同时保留了历史训练结果。

本文第 2 部分介绍了超球支持向量机理论,第 3 部分详细阐述了基于超球支持向量机的类增量学习算法,第 4 部分给出了在 Reuters 21578 标准语料库上的实验结果,最后得出结论。

2 超球支持向量机

设给定一类训练样本集 $\{x_i\}_{i=1}^{l}$ 和核函数 $K(x_i,x_j)$,其中,$x_i \in R^n$,K 对应某特征空间 Z 中的内积,即 $K(x_i,x_j)=\langle g(x_i),g(x_j)\rangle$,变换 $g:X \mapsto Z$ 是将样本从输入空间映射到特征空间。寻找特征空间的一个超球(a,R),其中 a 为球心,R 为球半径。超球应尽量包围样本的大部分映射,同时半径 R 应尽可能的小。当不存在偏远点时,则寻找一个能够包

*)基金项目:国家自然科学基金项目(No. 60603023);国家基础研究重大项目(973)研究专项(No. 2001CCA00700)。秦玉平 博士研究生,教授,主要研究领域为机器学习;李祥纳 硕士研究生,主要研究领域为机器学习;王秀坤 博导,教授,主要研究领域为数据库系统;王春立 博士,教授,主要研究领域为模式识别。

围所有样本映射的最小超球；当存在偏远点时，可以允许一部分样本的映射在超球的外面，则寻找一个能够包围大多数样本映射的最小超球。当不知道是否含有偏远点时，通过引入一个非负松弛变量 ξ_i，$i=1,2,\cdots,l$，允许一部分样本的映射位于超球的外面。采用与寻找最优分类面类似的方法，通过求解带约束条件式(2)和式(3)的优化问题式(1)得到最小包围球[11-13]。

$$\min F(R,a,\xi_i)=R^2+\frac{1}{vl}\sum_i\xi_i \tag{1}$$

$$s.t.\ \|g(x_i)-a\|^2\leqslant R^2+\xi_i \tag{2}$$

$$\xi_i\geqslant 0\quad i=1,\cdots,l \tag{3}$$

其中，$0<v\leqslant 1$，用来控制超球半径与它所能包围的样本数目之间的折衷。v 越小，惩罚也越大，对允许超球外面存在样本的约束程度也就越大。

为了求解上述优化问题，定义如下的 Lagrange 函数：

$$L(R,a,\beta,\gamma,\xi_i)=R^2+\frac{1}{vl}\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\beta_i\{R^2+\xi_i-\|g(x_i)-a\|^2\}-\sum_{i=1}^{l}\gamma_i\xi_i \tag{4}$$

其中 $\beta_i\geqslant 0$ 为该类样本集的 Lagrange 系数。

求解式(4)的最小值，可以令该泛函对 R,a 及 ξ_i 求偏导，并令导数等于 0，得到

$$\frac{\partial L}{\partial R}=2R\left(1-\sum_{i=1}^{l}\beta_i\right)=0\Rightarrow\sum_{i=1}^{l}\beta_i=1 \tag{5}$$

$$\frac{\partial L}{\partial \xi_i}=\frac{1}{vl}-\beta_i-\gamma_i=0\Rightarrow 0\leqslant\beta_i\leqslant\frac{1}{vl}\quad i=1,\cdots,l \tag{6}$$

$$\frac{\partial L}{\partial a}=-\sum_{i=1}^{l}2\beta_i(g(x_i)-a)=0\Rightarrow a=\sum_{i=1}^{l}\beta_i g(x_i)\quad i=1,\cdots,l \tag{7}$$

将约束条件式(5)、式(6)及式(7)代入式(4)中，并进行合并整理，得到

$$\max L(\beta_i)=L(R,a,\beta,\gamma,\xi_i)=\sum_{i=1}^{l}\beta_i K(x_i,x_i)-\sum_{i,j=1}^{l}\beta_i\beta_j K(x_i,x_j) \tag{8}$$

$$s.t.\ \sum_i\beta_i=1 \tag{9}$$

$$0\leqslant\beta_i\leqslant\frac{1}{vl}\ i=1,2,\cdots,l \tag{10}$$

由式(7)可知，最小超球球心为带权系数 β_i 的线性加权组合：

$$a=\sum_i\beta_i g(x_i) \tag{11}$$

当 $\beta_i>0$ 时，对应的样本称为支持向量。

当 $0<\beta_i\leqslant\frac{1}{vl}$ 时，对应的样本位于超球附近，任选其中一个该类样本 x 与超球球心之间的距离可确定超球半径。

$$R^2=\|g(x)-a\|^2=K(x,x)-2\sum_i\beta_i K(x,x_i)+\sum_{i,j}\beta_i\beta_j K(x_i,x_j) \tag{12}$$

当 $\beta_i=\frac{1}{vl}$ 时，对应的样本位于超球外面，称为野值或含噪声的样本。

3 超球支持向量机类增量学习算法

设给定样本集 $A=\{x_i,E_i\}_{i=1}^{l}$ 和核函数 $K(x_i,x_j)$，其中，$x_i\in R^n$，$E_i\in\{1,2,3,\cdots,N\}$，N 是样本集 A 中的类别数。K 对应某特征空间 Z 中的内积，即 $K(x_i^m,x_j^m)=\langle g(x_i^m),g(x_j^m)\rangle$，变换 $g:X\mapsto Z$ 是将样本从输入空间映射到特征空间。

设 A^m 为 A 中属于第 m 类的样本子集，其中，$m=1,2,\cdots,N$。对于每一类样本 A^m，利用超球支持向量机，在特征空间寻找一个超球 (a_m,R_m)，其中，a_m 是该类超球的球心，R_m 为超球的半径。

若有一类新类别样本集 A^{N+1} 加入，则利用超球支持向量机在特征空间确定其映射对应的超球 (a_{N+1},R_{N+1})，完成一次类增量学习。

对待分类样本 x，根据式(13)计算它的映射到每个超球球心 a_m 的距离 $d_m(x)$，$m=1,2,\cdots,N$。根据 $d_m(x)$ 来判断 x 是否属于该超球所包含的类。

$$[d_m(x)]^2=\|g(x)-a_m\|^2=\|g(x)-\sum_i\beta_i^m g(x_i^m)\|^2=K(x,x)+\sum_{i,j}\beta_i^m\beta_j^m K(x_i^m,x_j^m)-2\sum_i\beta_i^m K(x,x_i^m) \tag{13}$$

若对所有的 $d_m(x)$，都有 $d_m(x)>R_m$，根据式(14)计算待分类样本 x 属于第 m 类的隶属度。

$$r_m=\frac{R_m}{d_m(x)} \tag{14}$$

根据式(15)确定待分类样本 x 所属类别。

$$r=\arg\max_m r_m \tag{15}$$

若存在两个或两个以上的 $d_m(x)$，使得 $d_m(x)\leqslant R_m$，对满足此条件的 $d_m(x)$，根据式(16)计算待分类样本 x 属于第 m 类的隶属度。

$$r_m=1-\frac{d_m(x)}{R_m} \tag{16}$$

根据式(15)确定待分类样本 x 所属类别。

待分类样本 x 的分类过程具体描述如下：

步骤 1　根据式(13)计算 $d_m(x)$，$m=1,2,\cdots,N$；

步骤 2　若存在唯一的 r，使得 $d_r(x)\leqslant R_r$ 则 x 属于第 r 类，转步骤 5，否则转步骤转 3；

步骤 3　若对所有的 $d_m(x)$，$d_m(x)>R_m$ 都成立，先根据公式(14)计算样本 x 属于每个类别的隶属度 r_m，然后根据式(15)计算 x 所属的类别 r，转步骤 5，否则转步骤 4；

步骤 4　对每个 $d_m(x)$，若 $d_m(x)\leqslant R_m$，根据式(16)计算样本 x 属于该类的隶属度，然后根据式(15)计算 x 所属的类别 r，转步骤 5；

步骤 5　分类结束。

4 实验结果及分析

本文使用标准数据集 Reuters 21578，从中选取 5 类共 896 篇文本进行实验分析。使用其中的 598 篇作为训练样本，其余的 298 篇作为测试样本。将文本数据经过预处理后形成高维词空间向量，采用信息增益的方法来进行特征降维，向量中每个词的权重根据 tf-idf 公式计算。

表 1　训练语料和测试语料

类别	wheat	corn	coffee	soybean	cocoa
训练集规模	204	168	97	79	50
测试集规模	101	84	48	40	25

实验环境为 CPU Pentium 1.6G，内存 512M，操作系统 Windows XP。CIL 算法和 HS-CIL 算法都采用线性核函数，系统参数 $C=10$，$v=0.01$。算法实现参考了 Chang 和 Lin 所开发的 libsvm[14]，并在此基础上进行了相应的修改。

实验中采用通用的准确率、召回率和 F_1 值作为评价指标。

$$准确率=\frac{分类正确的样本数}{实际分类的样本数}$$

$$召回率=\frac{分类正确的样本数}{应有的样本数}$$

$$F_1=\frac{准确率\times召回率\times2}{准确率+召回率}$$

其中，每一类的准确率、召回率和 F_1 值称为微平均；所有类的准确率、召回率和 F_1 值称为宏平均。

实验中，初始样本集含有两类（第 1 类为 wheat，第 2 类为 corn），第 1 次增量为第 3 类（coffee）样本，第 2 次增量为第 4 类（soybean）样本，第 3 次增量为第 5 类（cocoa）样本。表 2 给出了本文算法和 CIL 算法的宏平均准确率、宏平均召回率和宏平均 F_1 值的比较。表 3 给出了本文算法和 CIL 算法训练时间和分类时间的比较。

表 2　HS-CIL 算法和 CIL 算法的宏平均准确率、宏平均召回率和宏平均 F_1 值的比较

学习过程	算法	宏平均准确率(%)	宏平均召回率(%)	宏平均 F1 值(%)
初始样本集	IL 算法	77.85	75.61	75.84
	HS-CIL 算法	72.76	72.84	72.79
第 1 次增量	CIL 算法	83.62	81.32	81.90
	HS-CIL 算法	79.75	78.09	78.85
第 2 次增量	CIL 算法	73.86	71.90	72.34
	HS-CIL 算法	74.97	71.53	73.21
第 3 次增量	CIL 算法	76.07	73.52	74.30
	HS-CIL 算法	76.64	73.47	75.02

表 3　HS-CIL 算法和 CIL 算法的训练时间和分类时间的比较

学习过程	算法	训练时间(ms)	分类时间(ms)
初始样本集	CIL 算法	265	126
	HS-CIL 算法	110	94
第 1 次增量	CIL 算法	218	343
	HS-CIL 算法	15	140
第 2 次增量	CIL 算法	328	502
	HS-CIL 算法	32	157
第 3 次增量	CIL 算法	203	891
	HS-CIL 算法	31	204

从实验结果可以看出，随着增量学习的不断进行，基于超球支持向量机的类增量学习算法与 CIL 算法相比，大大节省了训练时间，同时大大提高了分类速度，而分类性能基本一致。另外可以看出，当处理类别数多、样本数目大的数据集时，基于超球支持向量机的类增量学习算法的训练速度、分类速度和分类性能的优势表现得更加明显。

结束语　针对类增量学习，提出了一种基于超球支持向量机的类增量学习算法（HS-CIL）。增量学习过程中，只对新增类样本进行训练，就可将新类别知识加入到原有分类模型中，避免了由于所有样本都参加训练而导致的计算复杂度过高，避免了由于训练样本数量的不均衡而导致的分类精度降低，同时保留了历史训练结果。分类过程中，通过计算待分类样本到各超球球心的距离判定其所属类别，方法简单快捷，避免了由于过多的分类器计算而导致的分类速度下降。实验结果证明，该算法不仅具有较快的训练速度，而且具有较快的分类速度和较高的分类精度，对于类别数较多、数据集较大的情况，效果更加明显，是一种实用的类增量学习算法。

参 考 文 献

[1] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer, 1995

[2] Syed N, Liu H, Sung K. Incremental learning with support vector machines[A]. IJCAI[C]. Sweden: Stockholm, 1999

[3] Domeniconi C, Gunopulos D. Incremental support vector machine construction[A]. ICDM[C]. USA: California, 2001

[4] Cauwenberghs G, Poggio T. Incremental and Decremental Support Vector Machine[J]. Machine Learning, 2001, 44(13): 409-415

[5] Zhang Jinpei, Li Zhongwei, Yang Jing. A Divisional Incremental Training Algorithm of Support Vector Machine[A]// Proceeding of the IEEE International Conference on Mechatronics and Automation[C]. Canada, 2005, 8: 853-855

[6] 孔锐，张冰. 一种快速支持向量机增量学习算法[J]. 控制与决策，2005，20(10)：1129-1132

[7] 萧嵘，王继成，孙正兴，等. 一种 SVM 增量学习算法 α-ISVM [J]. 软件学报，2001，12(12)：1818-1824

[8] 张曦煌，须文波. 基于增量学习的超球支持向量机设计[J]. 计算机工程与应用，2006，42(13)：66-68，76

[9] 朱美琳，杨 佩. 基于支持向量机的多分类增量学习算法[J]. 计算机工程，2006，32(17)：77-79

[10] Zhang Bofeng, Su Jinshu, Xu Xin. A Class-incremental learning method for Multi-class Support Vector Machines in text classification[A]// Proceedings of the Fifth International Conference on Machine Learning and Cybernetics[C]. Dalian, 2006: 13-16

[11] 张翔，肖小玲，徐光祐. 基于样本之间紧密度的模糊支持向量机方法[J]. 软件学报，2006，17(5)：951-958

[12] 唐发明，王仲东，陈绵云. 支持向量机多类分类算法研究[J]. 控制与决策，2005，20(7)：746-749

[13] 张永，迟忠先，闫德勤. 一 种新的模糊补偿多类支持向量机[J]. 计算机科学，2006，33(12)：152-155

[14] Chang Chinchang, Lin Chihjen. LIBSVM: a Library for Support Vector Machines [J/OL]. http://www.csie.ntu.tw/~cjlin/lib-svm. 2005