DOI: 10.19650/j.cnki.cjsi.2005.s2.136
第 26卷第 8期增刊 仪 器 仪 表 学 报 2005年 8月

基于模糊 K近邻决策的柔性 SVM分类算法*

胡正平
(燕山大学通信电子工程系 秦皇岛 066004)

摘要 当海量样本之间相互混迭时,支持向量数目急剧增加,导致训练难度增大的同时 SVM分类器性能明显下降。针对该问题,在此构造模糊 KNN决策与支持向量机相结合的新的柔性 SVM分类器。它先建立所有训练样本的类间最近邻距离,根据各个训练数据的类间最近邻距离进行升序排列;然后根据模糊 k近邻分析结果对训练样本集进行修剪,在剩余空间中选择合适规模的样本子空间进行 SVM训练。在分类阶段,首先计算待识别样本和 SVM超平面的距离,如果距离大于某一设定门限,直接利用 SVM进行分类,否则带入到所有支持向量与修剪样本合成的模糊 KNN分类器中进行分类判决。对比实验结果表明,提出的算法无论是训练速度还是分类精度都远远好于单独的 SVM分类器。

关键词 支持向量机 模糊 K近邻分类器 最近邻

The Algorithm of Flexible SVM Classifier Based on Fuzzy KNN Analysis

Hu Zhengping
(*Department of Electronic Engineering, Yanshan University, Qinhuangdao 066004, China*)

Abstract some samples intermixed in another class seriously will result in the number of support vector will increase greatly and the performance of training and classification will become worse. To solve this problem, a novel fexlible SVM method based on fuzzy K nearest neighbor analysis (FKNN-FSVM) is proposed. Firstly a support vector interclass distance is defined and a sort process is presented. then the training samples are pruned, a sample is reserved or trimed according to the distribution of its k nearest neighbor, then the new subspace set is trained with SVM to obtain a classifier. In the classification phase, if the distance from the test sample to the optimal hyperplane is greater than the given threshold, the test sample would be classified by SVM; otherwise, the fuzzy KNN would be used. Experimental results show that this algorithm performs better than sole SVM in aspects of training speed and accuracy of classification.

Key words Support vector machine Fuzzy K nearest neighbor classifier Nearest neighbor

1 引 言

支持向量机建立在统计学习理论的 VC维概念以及结构风险最小原理的基础上[1],在逼近与分类应用领域取得了极大的成功,成为近年研究的热点。SVC是利用靠近边界的少数向量来构造最大间隔的分类超平面,当存在大规模训练集且样本之间存在相互混迭时,支持向量数目急剧增加,导致训练难度增大的同时 SVC分类器性能明显下降。针对这个问题,一些学者提出了不少解决问题的思路。W. J. Hu[2]等提出了加速分解的鲁棒支持向量机算法,通过引入归一化中心距离这一新的松弛变量,通过对松弛变量参数的控制,可以达到从不同范围选取支持向量的目的。文献[3,4]提出了 NN-SVM解决大规模支持向量机的学习问题,通过修剪训练样本的混迭样本,达到提高 SVM分类器的训练速度。文献[5]提出 RSVM(reduced Support Vector Machines),该方法通过随机选择训练样本子集,减少训练规模,提高训练速度。文献[6]提出了利用支持向量几何特征选择训练子集,进而加快训练速度

* 国家自然科学基金(60272073)资助项目

的算法。

这里提出的基于模糊 K近邻决策的柔性支持向量机分类算法是对 SVM 算法的扩展,它充分利用了支持向量分布的几何特征。

2 柔性支持向量机(FSVM)算法

构建的 FSVM原理框图分别如图 1图 2所示,该算法包括改进的训练算法和综合的分类算法两部分。

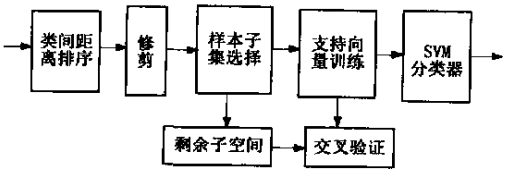

图 1 柔性支持向量机训练模型

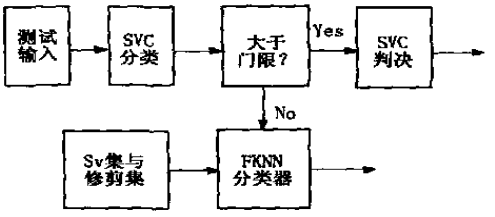

图 2 柔性支持向量机分类模型

2.1 训练数据最近邻类间距离排序

从图 3可以看出,由于支持向量大都集中在超平面附近并且它们相互之间比较接近,因此可以采取训练数据点的最小类间距离作为可能属于支持向量的测度。假设有一正样本(+ 1类)点 x_i,它到负样本(- 1类)最小的距离定义为:

$$D_i = D(x_i, y_j) = \min_{y_j \in \{-\text{类}\}} \| \phi(x_i) - \phi(y_j) \|$$
$$= \min_{y_j \in \{-\text{类}\}} (K(x_i, x_i) + K(y_j, y_j) - 2K(x_i, y_j))^{1/2} \quad (1)$$

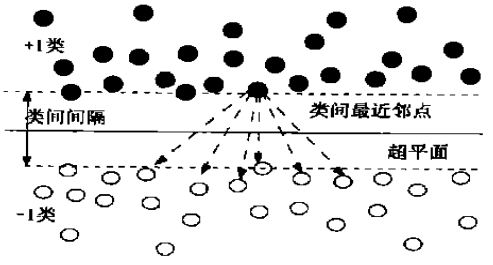

图 3 最近邻类间距离示意图

同样的方法,可以定义负样本(- 1类)点 y_j 到正样本(+ 1类)样本的距离。这里 x_i, i= 1, 2,…, n^+ 是类 + 1中的训练样本, y_j, j= 1, 2,… n^- 是类 - 1中的训练样本, K(·)为核函数。利用训练样本类间最近邻距离测度,分别对两类训练数据的进行升序排序。

2.2 基于 FKNN的训练数据修剪

首先找出每一个点 x的 k个近邻,然后对每一个近邻点,如果近邻点与该点类别相同,则标记 $L_i(x)$= 0,否则记为 1,设 $d_i(x)$, i= 1,…, k为 i近邻的欧氏距离。则模糊 k近邻混迭度定义为:

$$u(x) = \frac{\sum_{i=1}^{k} d_i^{-1}(x) \cdot L_i(x)}{\sum_{i=1}^{k} d_i^{-1}(x)} \quad (2)$$

根据上面定义的模糊 k近邻分析,结合类间最近邻距离排序结果,可以对训练数据中混迭严重的样本进行修剪。

(1)按照最近邻类间距离分别对两类样本进行升序排列,按照优先顺序选择一部分训练数据,计算其模糊 k近邻测度。

(2)根据每个训练数据的模糊 k近邻分析,设定一门限 β,如果大于或等于门限 β,则删除该点;否则,保留之。

(3)在修剪后的训练样本空间中,重新计算的类间最近邻距离,再选择合适的训练子集,用于 SVM的训练。

经过上述步骤之后,就可以得到修剪后的 SVM训练集,然后利用 SVM快速算法进行训练。

2.3 FSVM综合互补分类策略

将 FKNN与 SVM结合提高分类精度。FKNN-SVM综合互补分类策略如下:设 T为测试集, T_{SV}为支持向量集, T_d为被修剪样本集合。

(1)对于给定测试样本,首先计算它到支持向量机超平面的距离 d(x),如果 d(x)≥ th,则支持向量分类器直接输出。

$$d(x) = \sum_i y_i a_i K(x_i, x) - b \quad (3)$$

(2)如果 d(x) < th,代入 FKNN分类器,传递 x, T_{SV}, T_d, k 这里,在 FKNN分类器中,也可以采用原始空间的欧氏距离公式。

3 实　　验

为了验证文中提出的算法的性能,进行了两组实验。

(1)高斯分布样本点分类实验　随机产生了 2类三维样本的数据点,每类样本数目 15000点,满足多元正态分布,分布参数如表 1所示。

表 1 合成样本点参数分布

类别	均值 1	均值 2	均值 3	偏差
1	0.2	1.0	1.5	1
2	1.0	1.6	0.4	1

实验时,每一类别中选择 8000点作为训练样本,7000点作为测试样本。实验中采用高斯核函数(0.5),错误惩罚平衡因子 C等于 5 对比实验结果如表 2所示。

表 2 对比实验结果 1

方法	分类率(%)	Sv数	训练时间(s)
KNN修剪-SMO	71.76	501	3672
FKNN-SVM k= 1	73.17	467	1009
FKNN-SVM k= 3,β= 0.25	74.33	473	1028
FKNN-SVM k= 5,β= 0.2	74.51	479	1061

(2)MNIST手写数字识别实验

利用 MNIST手写数字数据库进行仿真实验,从60000个训练样本库选择了 5000个训练样本(“3”和“5”),从 10000个测试样本库中顺序选择 1000个作为测试样本,采用量化 PCA投影矢量作为训练特征,对比实验结果如表 3所示。

表 3 对比实验结果 2

方法	分类率(%)	SV数	训练时间(s)
KNN修剪 SMO	98.48	109	2016
FKNN-SVM k= 1	98.96	119	949
FKNN-SVM k= 3,β= 0.3	99.27	114	792
FKNN-SVM k= 5,β= 0.25	99.30	105	605
FKNN-SVM k= 7,β= 0.2	99.21	98	573

从上面的对比实验可以看出,文中提出的方法无论对于海量合成数据还是真实数据在保持分类精度的情况下,加快了 FSVM训练算法的速度。

4 结 论

提出了柔性支持向量机分类器的方法具有下面几个方面的优点:(1)根据各个训练数据的类间距离进行排序操作,便于选择合适的样本子空间,加快 SVM训练速度。(2)利用 FKNN进行混迭数据的修剪,通过修剪操作,可以减少支持向量的数目,加快分类的速度。(3)通过 FKNN-SVM综合互补分类策略,可以提高分类精度。研究结果对于解决 SVM在海量样本模式识别中的应用具有一定的作用。

参考文献

1 张学工.关于统计学习理论与支持向量机.自动化学报,2000,26(1):32~42.

2 W. J. Hu, Q. Song. An accelerated decomposition algorithm for robust support vector machines. IEEE Transactions on Circuits and Systems-Ⅱ: Express Briefs, 2004, 51(5): 234~240.

3 Li Hong-Lian, Wang Chun-hua, Yuan Bao-zong. An improved SVM: NN-SVM. Chinese Journal of Electronics, 2004, 13(2): 321~324.

4 Li Hong-Lian, Wang Chun-hua, Yuan Bao-zong. A learning strategy of SVM used to large training set. Chinese Journal of Computer, 2004, 27(5): 715~719.

5 Lin Kuan-Ming, Lin Chih-Jen. A study of reduced support vector machines. IEEE transactions on Neural Networks, 2003, 14(6): 1449~1459.

6 李青,焦李成,周伟达.基于向量投影的支持向量预选取[J].计算机学报,2005,28(2):145~151.