第34卷 第19期 计 算 机 工 程 2008年10月
Vol.34 No.19 Computer Engineering October 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)19—0206—03 文献标识码：A 中图分类号：TP301.6

基于搜索机制密度聚类的支持向量预选取算法

叶 菲，罗景青
(解放军电子工程学院信息系，合肥 230037)

摘 要：支持向量机在解决小样本、非线性及高维模式识别问题中具有许多特有的优势，但支持向量的选择过程复杂。该文利用聚类技术的特殊性能，提出基于搜索机制的密度聚类算法，该算法通过一种简单的搜索策略可将密度高于一定限度的对象聚为一类。将该算法用于支持向量的预选取，可减少训练样本数目，提高支持向量机的训练速度。从仿真实验可以看出，通过基于搜索机制密度聚类的支持向量预选取，训练样本数目可减少2/3以上，线性可分的数据训练速度可加快12倍左右，非线性可分的数据训练速度可加快5倍左右。
关键词：搜索机制；支持向量机；预选取

Pre-extracting Algorithm of Support Vector Based on Search Mechanism Density Clustering

YE Fei, LUO Jing-qing
(Department of Information, PLA Electronic Engineering Institute, Hefei 230037)

【Abstract】 Support Vector Machine(SVM) presents excellent performance to solve the problems with small sample, nonlinear and the problems of high-dimension pattern recognition, but the process of selecting support vector is quite complicated. Therefore a density clustering algorithm based on search is put forward. Through a sample search strategy the algorithm can cluster the object that its density is over certain threshold to one class, and the application of it to pre-extracting support vector can reduce the number of training samples and improve the training speed of SVM. From the simulation experiments, it can be found that through pre-extracting support vector based on search density clustering algorithm, the number of training sample can reduce 2/3, and the training speed can quicken 12 times for linear separable data and quicken 5 times for nonlinear separable data.
【Key words】 search mechanism; Support Vector Machine(SVM); pre-extracting

1 概述

支持向量机(Support Vector Machine, SVM)是基于统计学习理论而提出的一种小样本学习方法，具有很强的推广能力。从具体的运行过程来看，支持向量机的训练就是求解一个线性凸二次规划(Quadratic Programming, QP)问题。对于小规模的QP问题，经典最优化算法如牛顿法就可以较好地解决。但是当样本数量非常多时，训练就要花费大量的时间，这成为制约支持向量机应用于实际的主要问题。

如果能在不影响支持向量机性能的前提下，预先从已分类的样本中选取一部分作为训练样本，则不是将所有已分类样本作为训练样本，则可以减少训练样本的数目，加快支持向量机的训练速度。本文利用聚类分析技术的特殊性能，提出基于搜索机制的密度聚类算法，该算法通过一种简单的搜索策略，将密度高于一定限度的对象聚为一类。用该算法选取出处于分类边界的样本作为训练样本，以达到减少训练样本数量的目的。

2 支持向量机原理

支持向量机是一种新的机器学习技术[1-2]。基于统计学习理论的坚实基础，SVM有着很强的学习能力和较好的泛化性能。SVM分类方法是从线性可分情况下的最优分类面提出的。设线性可分样本集为 $(\boldsymbol{x}_i, y_i)$，$i=1,2,\cdots,n$，$\boldsymbol{x}_i \in R^d$，$y_i \in \{-1,+1\}$ 是类标号。d维空间中线性判别函数的一般形式为 $g(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\bullet\boldsymbol{x}+b$，分类面方程为 $\boldsymbol{w}^{\mathrm{T}}\bullet\boldsymbol{x}+b=0$。对判别函数进行归一化，使离分类面最近的样本的 $|g(\boldsymbol{x})|=1$，可得到分类间隔 $2/\|\boldsymbol{w}\|$，因此要求分类间隔最大等价于使 $\|\boldsymbol{w}\|$ 最小，而要求分类面将所有样本正确分类，则需满足

$$y_i[(\boldsymbol{w}^{\mathrm{T}}\bullet\boldsymbol{x}_i)+b]-1\geq 0,\quad i=1,2,\cdots,n \tag{1}$$

因此，满足上述条件且使 $\|\boldsymbol{w}\|$ 最小的分类面就是最优分类面。使式(2)等号成立的那些样本，支撑了最优分类面，被称作支持向量。权值向量 $\boldsymbol{w}$ 最小化代价函数为

$$\Phi(\boldsymbol{w})=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$$

利用Lagrange乘子法，可求得

$$\boldsymbol{w}=\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i$$

其中，α_i 是辅助非负变量，称作Lagrange乘子。则最优分类函数为

$$f(\boldsymbol{x})=sgn\{\sum_{i=1}^{n}\alpha_i y_i(\boldsymbol{x}\bullet\boldsymbol{x})+b\} \tag{2}$$

用内积核 $K(\boldsymbol{x}_i,\boldsymbol{x})$ 替代最优分类函数中的点积，就相当于将原特征空间变换到了另一新的特征空间，判别函数式为

$$f(\boldsymbol{x})=sgn\{\sum_{i=1}^{n}\alpha_i y_i K(\boldsymbol{x}_i\bullet\boldsymbol{x})+b\} \tag{3}$$

作者简介：叶 菲 (1980－)，女，博士研究生，主研方向：智能信息处理，空间信息处理；罗景青，教授、博士生导师
收稿日期：2007-12-26 **E-mail：**yefeixyz@163.com

这就是支持向量机，此时的最优分类算法就成了支持向量机分类算法。

式(2)和式(3)中的系数 α_i 是QP问题的解。在求得的解中，每一个系数 α_i 对应着一个训练样本，同时，存在许多系数 α_i 严格等于 0，只有那些具有非零系数的样本才会影响结果。因此，分类超平面只与这些样本有关，这些样本即为支持向量，从而可以看出，支持向量机的训练只与支持向量有关，而与非支持向量无关。从几何上直观地看，支持向量就是两类样本的交遇区内，那些靠的最近的处于两类样本边界上的样本。定义边界样本为某一类模式中位于其边界上的向量。则支持向量的预选取就是从大量已分类样本中取出边界样本[3-4]。

3 基于搜索机制的密度聚类算法

典型的密度聚类算法[5]是基于密度连接区域的密度聚类算法(Density-Based Spatial Clustering of Application with Noise, DBSCAN)，它需要计算每个点一定区域内包含对象的数目，根据一个密度阈值来控制簇的增长[6]。而本文基于搜索机制的密度聚类算法与此不同，它不需要计算每个对象一定区域内包含的对象个数，而是将待聚类对象集中某点作为一类，并从该点出发，通过距离参数搜索可能加入该类的数据点，而判别它们能否加入则是通过一个密度参数进行。在描述具体算法之前，先给出本算法中距离和密度的定义。

定义 1 设样本集为 $X=\{\boldsymbol{x}_i\}$，$i=1,2,\cdots,K$。$X\in R^n$ 是 n 维欧式空间的点集。距离函数为

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=1-\exp(-\sum_{k=1}^{n}(\boldsymbol{x}_i^k-\boldsymbol{x}_j^k)^2/2\sigma^2),\ \boldsymbol{x}_i,\boldsymbol{x}_j\in X$$

其中，σ 表示 RBF 函数的宽度参数。这样的距离函数是一个严格单调增加函数，具有正定性，对称性。且对 $\forall \boldsymbol{x}_i,\boldsymbol{x}_j\in X$，当 $\|(\boldsymbol{x}_i-\boldsymbol{x}_j)\|\to\infty$，有 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)\to 1$。

定义 2 假设样本集为 $X=\{\boldsymbol{x}_i\}$，$i=1,2,\cdots,K$，距离参数为 d_λ，则对于一个新增样本 $\boldsymbol{x}_k$，如果 X 中存在 m 个样本 $\boldsymbol{x}_j(j=1,2,\cdots,m)$ 和 $\boldsymbol{x}_k$ 之间的距离满足 $d(\boldsymbol{x}_j,\boldsymbol{x}_k)\le d_\lambda$，则称 $\boldsymbol{x}_k$ 基于集合 X 的密度为 m。

采用高斯型函数基于密度的聚类算法步骤如下：

(1)初始化距离参数 d_λ 和密度参数 $\lambda(0<\lambda\le 1)$，并从待聚类的样本集中任取一样本 $\boldsymbol{x}_i$ 作为起始样本$\boldsymbol{x}_0$。

(2)令集合 $A=\{\boldsymbol{x}_0\}$，计算 $\boldsymbol{x}_0$ 与其它样本 $\boldsymbol{x}_j$ 的距离 $d(\boldsymbol{x}_0,\boldsymbol{x}_j)$，如果 $d(\boldsymbol{x}_0,\boldsymbol{x}_j)\le d_\lambda$，则将 $\boldsymbol{x}_j$ 加入集合 A。如果不存在 $\boldsymbol{x}_j$ 与 $\boldsymbol{x}_0$ 的距离小于 d_λ，集合 $A=\{\boldsymbol{x}_0\}$ 即为聚出的一类，从剩余样本集中另取一样本作为起始样本 $\boldsymbol{x}_0$，重复步骤(2)。

(3)令集合 A 的样本个数为 m，从 A 中取一未作过起始样本的样本为起始样本 $\boldsymbol{x}_0$，计算 $\boldsymbol{x}_0$ 与其他不属于集合 A 的未聚类样本 $\boldsymbol{x}_k$ 的距离 $d(\boldsymbol{x}_0,\boldsymbol{x}_k)$，如果 $d(\boldsymbol{x}_0,\boldsymbol{x}_k)\le d_\lambda$，且样本 $\boldsymbol{x}_k$ 基于集合 A 的密度大于 λm，则将 $\boldsymbol{x}_k$ 加入集合 A。

(4)重复步骤(3)，直到 A 中所有样本都曾被作为起始样本为止，则 A 为聚出的一类。

(5)从剩余样本集中任取一样本作为起始样本 $\boldsymbol{x}_0$，回到步骤(2)，直到没有剩余样本为止，聚类结束。

(6)为了从不同尺度观察样本间的相异程度及类间的关系，可从步骤(1)开始，调整距离参数，重新聚类。

该聚类算法按照距离参数进行搜索，不需要给定期望类数，完全根据样本属性的性质进行聚类，并可通过密度参数 λ 来控制聚类的形状。

4 基于密度聚类的支持向量预选取

与支持向量机相对应，基于搜索机制密度聚类的支持向量预选取方法也分为线性的和非线性两种形式，分别用于线性和非线性支持向量机。

4.1 线性 SVM 的支持向量预选取

聚类分析是依据样本间关联的度量标准将其自动分成几个群组，且使同一个群组内的样本相似，而属于不同群组的样本相异的方法。因此，常用聚类算法是尽量将属于一类的样本归为一类，而支持向量预选取过程中的聚类是要尽量找出哪些不属于一类却容易被聚为一类的样本，即寻找边界样本，因为处于两类样本边界上的样本是最容易被错误分类的。因此，可以采用本文提出的聚类算法，通过调整距离参数的大小，从不同尺度观察聚类的结果，从而找出最易于被错误聚类的样本。在描述具体算法之前，给出一些定义。

定义 3 某一类样本的平均特征称为该类的中心 $\boldsymbol{m}$，已知样本向量组 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，其中，样本向量为 N 维，那么其中心 $\boldsymbol{m}$ 也是 N 维向量，为

$$\boldsymbol{m}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \tag{4}$$

定义 4 为了表示样本之间的相异程度，定义两个 N 维样本 $\boldsymbol{x}_1$，$\boldsymbol{x}_2$ 之间的样本距离为

$$d(\boldsymbol{x}_1,\boldsymbol{x}_2)=1-\exp(-\sum_{i=1}^{N}(\boldsymbol{x}_1^i-\boldsymbol{x}_2^i)^2/2\sigma^2) \tag{5}$$

定义 5 某一类样本到该类中心的距离称为中心距离。假设样本向量为 $\boldsymbol{x}$，类中心为 $\boldsymbol{m}$，中心距离为

$$d(\boldsymbol{x},\boldsymbol{m})=1-\exp(-\sum_{i=1}^{N}(\boldsymbol{x}^i-\boldsymbol{m}^i)^2/2\sigma^2) \tag{6}$$

定义 6 类中各样本中心距离的平均值称为该类的平均距离，为

$$p=\sum_{i=1}^{n}d(\boldsymbol{x}_i,\boldsymbol{m}) \tag{7}$$

定义 7 对属于 $\{-1,+1\}$ 两类的样本进行聚类，则聚类后某类中属于 $+1$ 类的样本有 N_{+1} 个，属于 -1 类的样本有 N_{-1} 个，定义该类的样本正确率为

$$r=\frac{\max(N_{+1},N_{-1})}{N_{+1}+N_{-1}} \tag{8}$$

则有 $0.5\le r\le 1$。

假设已分类样本集为 $(\boldsymbol{x}_i,y_i)$，$i=1,2,\cdots,n$，$\boldsymbol{x}_i\in R^d$，$y_i\in\{-1,+1\}$ 是类标号。则基于密度聚类的支持向量预选取过程如下：

(1)计算 $+1$ 类的平均距离 p_1 和 -1 类的平均距离 p_2。

(2)设距离参数 d_λ 为 $\min(p_1,p_2)$。

(3)用本文提出的聚类算法对样本集进行聚类。如果聚类后每类的样本正确率都为 1，则 d_λ 值是适当的。如存在某类的样本正确率小于 1，则表示 d_λ 取值稍大。减小 d_λ，使每类的样本正确率都为 1，此时类的数目可能大于 2。

(4)取步长 λ，以 λ 增大距离参数 d_λ，重新聚类。随着距离参数的增大，类的数目会随之减小，两类间的边界样本则会逐渐被聚为一类。此时必有一类或多类的样本正确率开始下降，这些样本正确率小于 1 的类中的样本就是边界样本。

(5)逐渐增大距离参数 d_λ，直到边界样本的数目满足训练样本数的要求为止。

4.2 非线性 SVM 的支持向量预选取

非线性可分的模式采用非线性映射 Φ 把输入空间映射到某一特征空间 H 中，令 $K(\bullet,\bullet)$ 为内积核，则样本向量的距离为

$$d^H(\boldsymbol{x}_1,\boldsymbol{x}_2)=1-\exp(-(K(\boldsymbol{x}_1,\boldsymbol{x}_1)-2K(\boldsymbol{x}_1,\boldsymbol{x}_2)+K(\boldsymbol{x}_2,\boldsymbol{x}_2))/2\sigma^2) \quad (9)$$

输入空间样本的中心经映射后得到的值不再是特征空间中样本的中心，特征空间样本的中心向量要在特征空间中求得，即

$$m_\Phi=\frac{1}{n}\sum_{i=1}^{n}\Phi(\boldsymbol{x}_i) \quad (10)$$

其中，n 是样本的个数。因为不知道映射 $\Phi(\boldsymbol{x})$ 的具体表达形式，样本的中心距离不能通过式(6)求出，而定义为

$$d^H(\boldsymbol{x},\boldsymbol{m}_\Phi)=1-\exp(-(K(\boldsymbol{x},\boldsymbol{x})-\frac{2}{n}\sum_{i=1}^{n}K(\boldsymbol{x},\boldsymbol{x}_i)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(\boldsymbol{x}_i,\boldsymbol{x}_j))/2\sigma^2) \quad (11)$$

同式(7)，类的平均距离可由样本中心距离求得，为

$$p=\sum_{i=1}^{n}d^H(\boldsymbol{x}_i,\boldsymbol{m}_\Phi)$$

基于上述定义，非线性 SVM 支持向量预选取算法的具体步骤与线性 SVM 支持向量预选取算法是相同的。

5 仿真分析

为了验证支持向量预选取算法的有效性，可以通过以下实验进行验证。

(1)线性可分的例子

随机产生两类线性可分的均匀分布数据，用星号和圈号表示，如图 1 所示。利用本文的密度聚类算法，聚出的各类用椭圆曲线表示，其中带阴影的类包含的样本数据即为边界样本，而处于分类边界上的边界样本就是支持向量。表 1 是没有经过支持向量预选取和经过支持向量预选取时，SVM 的分类结果比较，从表中可以看出，两种支撑向量机的分类效果是完全一致的，但经过支持向量的预选取可以减少运算量。

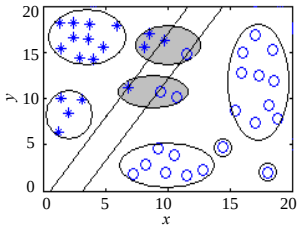

图 1 线性可分时的边界样本选取

表 1 线性可分时分类性能比较

算法	训练样本	边界样本	支持向量	训练时间/s	检验样本	识别率/(%)
未加预选取算法的 SVM	40	未选	4	15.832 9	100	100
增加预选取算法的 SVM	40	7	4	1.237 4	100	100

(2)非线性可分的例子

随机产生两类均匀分布的非线性可分数据，分别用星号和加号表示，如图 2 所示。采用高斯核函数：

$$K(x,y)=\exp(-\frac{(x-y)^2}{2\sigma^2})$$

取 $\sigma=0.1$。图中加圈号的样本是预选取的边界样本，加矩形的是训练得到的支持向量。表 2 是没有经过支持向量预选取和经过支持向量预选取时，SVM 的分类结果比较，从表中可以看出，两种支撑向量机的分类效果大致相同，但经过支持向量的预选取可以减少运算量。

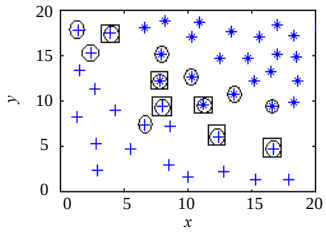

图 2 非线性可分时的边界样本选取

表 2 非线性可分时分类性能比较

算法	训练样本	边界样本	支持向量	训练时间/s	检验样本	识别率/(%)
未加预选取算法的 SVM	40	未选	8	22.451 6	100	94
增加预选取算法的 SVM	40	13	6	3.816 3	100	93

在有限先验知识和高维输入矢量的情况下，支持向量机具有较好的推广性，被应用于雷达信号分选与识别处理中。由于雷达信号密度已达到一秒钟几百万个脉冲数，训练数据异常庞大，使得训练速度非常缓慢。如果采用支持向量的预选取算法，可以大大减少雷达信号的训练样本数，加快训练速度，从而满足实际应用的需要。

6 结束语

本文通过对基于密度聚类算法的研究，提出了一种改进的基于搜索机制的密度聚类算法。用该聚类算法选取位于分类边界附近的样本作为训练样本，从而大大减少了训练样本的数目，提高支持向量机的训练速度。

参考文献

[1] Haykin S. 神经网络原理[M]. 北京: 机械工业出版社, 2003.

[2] 焦李成. 智能数据挖掘与知识发现[M]. 西安: 西安电子科技大学出版社, 2006

[3] 焦李成, 张 莉. 支撑矢量预选取的中心距离比值法[J]. 电子学报, 2001, 29(3): 383-386.

[4] 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): 145-152.

[5] 王劲波, 翁 伟. 数据挖掘中基于密度的聚类分析算法[J]. 统计与决策, 2005, 20(20): 139-141.

[6] 武方方, 赵银亮. 基于密度聚类的支持向量机分类算法[J]. 西安交通大学学报, 2005, 39(2): 1319-1322.