CNKI :50-1165/N. 20110917. 1359. 011

# 改进的双隶属度模糊支持向量机*

邬 啸, 魏 延, 吴 瑕
(重庆师范大学 计算机与信息科学学院, 重庆 401331)

**摘要**:针对传统的支持向量机(SVM)中存在对噪声和孤立点敏感,容易产生过拟合的问题,提出一种新的模糊隶属度函数设计方法——基于密度法的双隶属度模糊支持向量机方法(DM-FSVM)。该方法不仅考虑样本到类中心的距离,同时根据样本点到类中心的距离将样本分为两类,类中心附近样本点的隶属度由该样本点到类中心的距离确定,而对于远离类中心的样本点来说,其隶属度由邻域内同类与异类样本点数目的比值来确定。同时,针对模糊支持向量机普遍存在训练时间过长的难题,使用截集模糊 $C$-均值聚类的方法对训练样本进行聚类处理,以聚类中心作为新的样本进行训练。最后数值实验表明,与传统的支持向量机和以往的 FSVM 相比,有效地提高了分类速度和精度。

**关键词**:支持向量机;双隶属度;截集模糊 $C$-均值
**中图分类号**:TP181;TP311.131 **文献标志码**:A **文章编号**:1672-6693(2011)05-0049-04

支持向量机(Support vector machine,SVM)是 Vapnik 等人提出的一种基于统计学习理论的新的机器学习方法[1-6],由于其具有良好的性能和广泛的应用[7-8]而日益受到重视,形成近年来的研究热点。支持向量机的处理机制决定了其对训练样本内的噪音和孤立点非常敏感,对不是完全属于两类中的一类的样本分类正确率偏低。针对这些不足,Lin 等学者[9-10]通过对每一个样本引入一个模糊隶属度参数,对不同的样本采用不一样的惩罚系数,使得不同的样本在构造目标函数时有不一样的贡献,而构建了模糊支持向量机(Fuzzy support vector machine,FSVM)。其对噪声或野值点赋予非常小的权值,从而有效达到消除噪声或孤立点、提高分类精度的目的。

在模糊支持向量机中,隶属度函数的设计是整个算法的关键。它必须能够准确、客观地反映样本点的不确定性;具有优良的去噪声和孤立点的能力;同时还要控制隶属度函数算法的复杂度,以提高运算效率。文献[9]给出了一种基于类中心的隶属度函数设计方法,该方法简单易行,每个样本点的隶属度根据其到类中心的距离确定,但其中类半径的确定对噪声和孤立点很敏感。文献[10]提出了一种基于密度法的隶属度函数设计思想,各样本点的隶属度由其邻域内同类异类样本点的比例确定,但算法复杂度较高,计算机运算时间明显增加。

## 1 模糊支持向量机

模糊支持向量机可以减小甚至忽略非重要样本和噪声点对支持向量机学习的影响,从而提高分类精度[11]。在采用模糊支持向量机时,首先要对数据进行预处理:选择一个合适的隶属度函数,得到每一个样本 $x_i$ 的隶属度函数 $\mu(x_i)$。于是便可以得到新的模糊训练集:$\{x_1, y_1, \mu(x_1)\}$, $\{x_2, y_2, \mu(x_2)\}$,…, $\{x_n, y_n, \mu(x_n)\}$。每个训练点 $x_i \in \mathbf{R}^d$, $y_i \in \{-1, +1\}$, $0 \leqslant \mu(x_i) \leqslant 1$。其中 $\mu(x_i)$ 为训练点 $\{x_i, y_i, \mu(x_i)\}$ 的输出 $y_i = 1$(正类)或 $-1$(负类)的模糊隶属度($i = 1, 2, \cdots, n$)。

模糊隶属度指 $\mu(x_i)$ 是训练点 $\{x_i, y_i, \mu(x_i)\}$ 隶属于某一类的程度,而 $\xi_i$ 是对错分程度的度量,所以就用 $\mu(x_i)\xi_i$($i = 1, 2, \cdots, n$)衡量对于重要程度不同的变量错分程度。由此得到最优分类超平面为下面目标函数的最优解。

$$\begin{aligned} &\min\ \varphi(\omega, \xi) = 0.5\|\omega\|^2 + C\sum_{i=1}^{n}\mu(x_i)\xi_i \\ &\text{s.t.}\quad y_i[\omega'\varphi(x_i) + b] \geqslant 1 - \xi_i \\ &\xi_i \geqslant 0 (i = 1, 2, \cdots, n) \end{aligned} \tag{1}$$

* 收稿日期:2011-05-24 修回日期:2011-07-08 网络出版时间:2011-09-17 13:59:00
资助项目:重庆市教委科学技术研究项目(No. KJ090823)
作者简介:邬啸,男,硕士研究生,研究方向为机器学习与智能计算;通讯作者:魏延,E-mail:weiyancq@126.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20110917.1359.201105.49_011.html

其中:惩罚因子 $C$ 为常量,$\omega$ 表示线性分类函数的权系数,$\xi=(\xi_1,\xi_2,\cdots,\xi_n)^{\mathrm{T}}$,$\varphi$ 将 $x_i$ 从 $\mathbf{R}^d$ 映射到高维特征空间。而相应的最优分类超平面的判别式为

$$f(x)=\operatorname{sgn}\Big[\sum_{i=1}^{n}\alpha_i^* y_i K(x_i,x)+b^*\Big]$$

$$0\leqslant\alpha_i^*\leqslant\xi_i C,i=1,2,\cdots,n \qquad (2)$$

其中 $K(x_i,x)$ 称为核函数,$K(x_i,x)$ 将高维特征空间的内积运算转化为低维模式空间上的一个简单的函数计算。

由(1)式可以看出,当 $\mu(x_i)$ 很小时,减少了 $\xi_i$ 在式中的影响,以至于将相应的 $x_i$ 可以视为不重要的样本。对于孤立点或噪音样本,如果能够使其 $\mu(x_i)$ 很小,则此样本对支持向量机的训练作用就大为降低,进而降低了它们对训练支持向量机的影响。由此可见模糊因子 $\mu(x_i)$ 的确定正是这种模糊支持向量机工作性能好坏的关键。

## 2 $C$-均值聚类

### 2.1 聚类算法

将抽象对象的集合分成相似的对象类的过程称为聚类[12-14]。聚类是一种非监督模式的识别问题,它是指按照某种相似性的度量,使相似的样本归为相同的类。目前已经广泛应用于许多领域,包括模式识别、数据分析和图像处理等。

$C$-均值算法以 $C$ 为参数,将 $n$ 个对象分为 $C$ 个簇,使簇内具有较高的相似度,而簇间的相似度则较低。相似度的计算是根据某个簇中的对象的平均值来确定。其核心思想是把 $n$ 个向量 $x_j(j=1,2,\cdots,n)$ 分为 $c$ 个组 $G_i(i=1,2,\cdots,c)$,并求每组的聚类中心,使得非相似性指标的目标函数达到最小。但是该算法对初始点非常敏感,常常对于不同的初始聚类中心会得到不一样的聚类结果。基于此,本文采用一种新的高效软聚类方法 — 截集模糊 $C$-均值聚类[15]对样本个数进行优化简约以提高运算速度。

### 2.2 截集模糊 $C$-均值聚类算法

模糊 $C$-均值聚类是一种比较高效的算法,它使一个样本以不同的隶属度属于所有的类。但是一般情况下,若一个样本属于某类的隶属度远远大于属于其他类的隶属度,可以用最大隶属度原则,将其划为该类,不必再考虑其他的情况,只有当属于各个类的隶属度都非常接近的样本,才将其与各类相联系。截集模糊 $C$-均值聚类是模糊 $C$-均值聚类的一个改进[16],可以加快聚类速度并且使聚类更加合理。

设数据集 $X\subset\mathbf{R}^{s\times n}$ 的模糊 $C$-划分 $U=[u_{ik}]_{c\times n}$,$1\leqslant i\leqslant c$,$1\leqslant k\leqslant n$ 及 $\lambda\in[0,1]$,令

$$U_{pk}=\max\{U_{ik},1\leqslant i\leqslant c\} \qquad (3)$$

$$w_{ik}=\begin{cases}1,U_{pk}\geqslant\lambda\text{ 并且 }i=p\\0,U_{pk}\geqslant\lambda\text{ 并且 }i\neq p\\u_{ik},U_{pk}<\lambda,\forall 1\leqslant i\leqslant c\end{cases} \qquad (4)$$

则 $W=[w_{ik}]_{c\times n}$ 即为数据集 $X\subset\mathbf{R}^{s\times n}$ 的 $\lambda$ 截集模糊 $C$-划分[17]。

根据 Bezdek 的 ISODATA 算法来设计截集模糊 $C$-均值聚类算法(S2FCM),实现步骤如下。

初始化指数因子 $m$,停止误差 $\varepsilon$,分类类数 $c$,截集因子 $\lambda=0.5+\frac{1}{2c}$。

第1步 在数据集 $X=\{x_1,x_2,\cdots,x_n\}$ 中,随机选择 $c$ 个数据作为初始聚类中心 $V=\{v_1,v_2,\cdots,v_c\}$[13];

第2步 计算 $x_k(1\leqslant k\leqslant n)$ 到聚类中心 $v_i(1\leqslant i\leqslant c)$ 的内积范数 $d_{ik}=\|x_k-v_i\|_\lambda^2$;

第3步 令 $d_{c+lk}=\min\{d_{ik},1\leqslant i\leqslant c\}(1\leqslant k\leqslant n)$,返回 $d_{c+lk}$ 值所对应的 $i$ 的值,记为 $s$,并计算 $p_k=\Big(\frac{1}{d_{c+lk}}\Big)^{\frac{1}{m-1}}\Big/\sum_{i=1}^{c}\Big(\frac{1}{d_{ik}}\Big)^{\frac{1}{m-1}}$。对每一个 $x_k(1\leqslant k\leqslant n)$ 进行如下处理:

若 $d_{c+lk}=0$ 或 $d_{c+lk}\neq 0$ 且 $p_k\geqslant\lambda$,则 $w_{ik}=\begin{cases}1,i=s\\0,i\neq s,\forall 1\leqslant i\leqslant c\end{cases}$;若 $d_{c+lk}\neq 0$ 且 $p_k<\lambda$,则 $w_{ik}=\Big(\frac{1}{d_{lk}}\Big)^{\frac{1}{m-1}}\Big/\sum_{i=1}^{c}\Big(\frac{1}{d_{ik}}\Big)^{\frac{1}{m-1}}$。

通过以上的计算即可得到 $W=[w_{ik}]_{c\times n}$;

第4步 取出不为零的 $w_{ik}$ 来计算 $v_i=\sum_{k=1}^{n}(w_{ik})^m x_k\Big/\sum_{k=1}^{n}(w_{ik})^m$,得到新的聚类中心 $V^{l+1}$;

第5步 若 $\|V^{l+1}-V^l\|^2\leqslant\varepsilon$,则算法终止,并且输出聚类中心 $V^{l+1}$,否则令 $l=l+1$,返回第2步。

## 3 改进的模糊支持向量机

基于上述理论,在本文中给出了一种新的隶属度函数设计方法 — 基于截集模糊 $C$-均值聚类和密度法的双隶属度模糊支持向量机方法(DM-FSVM)。首先,使用 $C$-均值聚类的方法对训练样本进行聚类处理,以聚类中心作为新的样本进行训练;其次,产生的新样本中,类中心附近样本点的隶属度由该样本点到类中心的距离确定,而对于远离类中心的样本点来说,其隶属度由邻域内同类与异类样本点数目的比值来确定。

3.1 定义

给定训练样本集 $T = \{x_1, y_1, \mu_1\}, \{x_2, y_2, \mu_2\}, \cdots, \{x_n, y_n, \mu_n\}$。样本点之间的距离为 $D(x_i, y_j) = \|x_i - y_j\|$。样本点的同类点密度、异类点密度分别为

$$\rho^+(x_i, R) = |\{x_j \mid D(x_j, x_i) \leqslant R, y_j = y_i\}|$$

$$\rho^-(x_i, R) = |\{x_j \mid D(x_j, x_i) \leqslant R, y_j \neq y_i\}|$$

其中 $R$ 为可调节的样本点邻域半径。

样本点的类中心：$O^+ = \dfrac{\sum_{y_i=1} x_i}{l^+}, O^- = \dfrac{\sum_{y_i=-1} x_i}{l^-}$。其中 $l^+$、$l^-$ 分别为正负类样本点的数目。

调节 $O^+$、$O^-$ 的邻域半径，使得

$$\rho(O^+) = |\{x_j \mid D(x_j, O^+) \leqslant R^+, y_j = 1\}| = \alpha\% l^+$$

$$\rho(O^-) = |\{x_j \mid D(x_j, O^-) \leqslant R^-, y_j = -1\}| = \alpha\% l^-$$

其中 $0 < \alpha < 100$，为可调节参数。

这样，由类中心和类半径组成的超球只覆盖了其中 $\alpha\%$ 的类样本点（$\alpha\%$ 的含义是通过调节 $\alpha$ 这个参数使得只将部分样本点包括进去，这样即可有效地去除野值点），大大降低了噪声和孤立点对类半径的影响。

3.2 模糊隶属度函数的确定

根据样本点到其类中心的距离将样本点分为两类，分别采用不同的隶属度设计方法。

3.2.1 类中心附近样本点的隶属度　类中心附近的点是指符合下面条件的样本：

$$D(x_i, O^+) \leqslant \beta R^+, y_i = +1$$

或

$$D(x_i, O^-) \leqslant \beta R^-, y_i = -1$$

其中 $0 < \beta \leqslant 1$ 为可调节参数。

则其隶属度函数为

$$\mu_i = \begin{cases} \dfrac{1}{1 + \dfrac{\gamma D(x_i - O^+)}{R^+}}, y_i = +1 \\ \dfrac{1}{1 + \dfrac{\gamma D(x_i - O^-)}{R^-}}, y_i = -1 \end{cases}$$

其中 $\gamma > 0$ 为可调节参数。

3.2.2　远离类中心的样本点隶属度　远离类中心的样本点是指符合下列条件的点：

$$D(x_i, O^+) > \beta R^+, y_i = +1$$

或

$$D(d_i, O^-) > \beta R^-, y_i = -1$$

其中 $0 < \beta < 1$ 为可调节参数。则其隶属度函数为 $\mu_i = \dfrac{\eta\rho^+(x_i, \theta)}{\rho^+(x_i, \theta) + \rho^-(x_i, \theta)}$，其中 $0 < \eta \leqslant 1, \theta > 0$ 为可调节参数。

如此，既准确地反映出不同样本点对支持向量的影响，又降低了算法的复杂度。

3.3 新的模糊支持向量机算法

步骤 1　将原始的样本数据首先按照 2.2 节中截集模糊 $C$-均值聚类算法进行一次简约，将产生的聚类中心作为新的样本数据；

步骤 2　根据 3.2 节中提出的模糊隶属度参数确定方法，计算出新样本中各点的隶属度；

步骤 3　训练支持向量机。

算法的计算时间消耗主要在聚类后的样本集中样本点之间的距离和对距离的扫描检索上，步骤 1 中基于截集模糊 $C$-均值聚类算法，当样本集 $T$ 中所有样本点聚类完毕后结束，步骤 2 中 $T$ 中样本总数为 $l$，故其时间复杂度是 $O(l^2)$。

## 4 仿真实验

通过对样本训练时间的对比，可以明显看出，聚类后的训练时间要少于标准 SVM 和文献[9]中提出的 FSVM 算法，以德国信用数据（German credit data）为例，该数据库有 1 000 个样本，+1 类样本 700 个，-1 类样本 300 个，每个样本有 24 个信用信息指标。从中随机抽取 667 个样本作为训练集，其余 333 个样本作为测试集。结果见表 1。

表 1　3 种算法的训练时间比较

| 数据集（样本个数）（训练/测试） | 算法 | 运行时间/s |
|---|---|---|
| German credit | SVM | 58.31 |
| data | FSVM | 97.52 |
| （667/333） | DM-FSVM | 36.56 |

实验用的数据集 Splice 和 Ijcnnl 均从 UCI 数据库下载，分别采用标准支持向量机算法（SVM），文献[9]中提出的 FSVM 算法，以及本文中提出的 FSVM（DM-FSVM）。核函数采用径向基函数，在其他参数都相同的情况下，分别运行 20 次，然后取其平均值，比较结果见表 2。

表 2　3 种算法的运行时间和精确度比较

| 数据集（样本个数）（训练/测试） | 算法 | 运行时间/s | 分类精度/% |
|---|---|---|---|
| Splice（1000/2175） | SVM | 2.92 | 79.86 |
| | FSVM | 4.06 | 83.52 |
| | DM-FSVM | 2.16 | 84.27 |
| Ijcnnl（49990/91701） | SVM | 118.97 | 86.53 |
| | FSVM | 152.56 | 86.97 |
| | DM-FSVM | 76.17 | 87.06 |

## 5 结束语

提出了基于截集模糊 $C$-均值聚类和密度法的双隶属度模糊支持向量机方法(DM-FSVM),通过先对原始数据进行有效聚类,以聚类中心作为样本,减少样本数量,再使用双隶属度的方法,对样本进行不同的处理,使得在分类速度和精度上都得到提高。通过仿真实验证明了该方法的有效性,从而大大提高了支持向量机的泛化能力。对于如何判定样本中的噪声或孤立点,是笔者未来的研究方向。

## 参考文献:


[1] Vapnik V. The nature of statistical learning theory [M]. New York: Springer, 1995.

[2] Cristianini N, Taylor S J. An introduction to support vector machines [M]. Cambridge: Cambridge University Press, 2000.

[3] 万辉,魏延. 一种改进的最小二乘支持向量机算法[J]. 重庆师范大学学报:自然科学版,2010,27(4):69-72.

[4] 汪洋,陈友利,刘军,等. 基于相似方向的二叉树支持向量机多类分类算法[J]. 四川师范大学学报:自然科学版,2008,31(6):762-765.

[5] 郑宏宇,贺瑞缠. 一种新的支持向量机光滑函数[J]. 重庆工学院学报:自然科学版,2009,23(12):134-137.

[6] 吴渝,向浩宇,刘群. 一种基于网格的最近邻 SVM 新算法[J]. 重庆邮电大学学报:自然科学版,2008,20(6):706-709.

[7] 张浩然,韩正之,李昌刚. 基于支持向量机的非线性系统辨识[J]. 系统仿真学报,2003,15(1):119-121.

[8] 张炤,张素,章琛曦,等. 基于支持向量机的概率密度估计方法[J]. 系统仿真学报,2005,17(10):2355-2357.

[9] Lin C F, Wang S D. Fuzzy support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471.

[10] 安金龙,王正欧,马振平. 基于密度法的模糊支持向量机[J]. 天津大学学报,2004,37(6):544-548.

[11] 边肇祺,张学工. 模式识别[M]. 北京:清华大学出版社,2007.

[12] 罗军生,李永忠. 基于模糊 $C$-均值聚类算法的入侵检测[J]. 计算机技术与发展,2008,18(1):178-180.

[13] 王伟,高亮. 一种基于模糊聚类的离散化方法[J]. 计算机技术与发展,2008,18(3):53-55.

[14] 吴瑛,王秋生. 模糊 $C$-均值聚类算法在 Web 使用挖掘上的应用研究[J]. 计算机技术与发展,2008,18(6):32-35.

[15] 裴继红,范九伦,谢维信. 一种新的高效软聚类方法:截集模糊 $C$-均值(S2FCM)聚类算法[J]. 电子学报,1998,26(2):83-86.

[16] Bezdek J C. Pattern recognition with fuzzy objective function algorithms [M]. New York: Plenum Press, 1981.

[17] Li M J, Ng M K, Cheung Y M, et al. Agglomerative fuzzy K-Means clustering algorithm with selection of number of clusters [J]. IEEE Trans Knowledge and Data Engineerning, 2008, 20: 1519-1534.


## Improved Double Memberships of Fuzzy Support Vector Machine


*WU Xiao, WEI Yan, WU Xia*

(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)



**Abstract**: In traditional support vector machine, there is an existence of noise and outlier sensitive, and often has an over-fitting problem. In this paper, a new membership of fuzzy support vector machine desigh method is proposed-double memberships of fuzzy support vector machine(DM-FSVM). Not only consider the distance of samples to center, but also divide the sample points into two types according to the distance of sample points to the center, near the clustering center sample 's membership determined by the distance from this sample to the clustering center, but for the other samples which are far away from clustering center, its membership is a ratio of samiler and disparate class points within its neighborhood. Meanwhile fuzzy vector machine widespread exists long training time probleml, using cut sets $C$-mean clustering method for training data to clustering process and treat the clustering center as new samples to training. Experimental results show the good performance of the present approach.



**Key words**: SVM; double memberships; $C$-mean clustering


(责任编辑 游中胜)