文章编号: 1001-0920(2008)09-1030-05

基于模糊一类支持向量机的核聚类算法

徐 磊, 赵光宙
(浙江大学 电气工程学院, 杭州 310027)

摘 要: 引进模糊概念替代距离拒绝尺度,定义具有支持向量特性的模糊隶属度函数,以描述训练点隶属于聚类集的程度. 惩罚了边缘点对聚类中心的贡献权重,从而抑制了聚类中心的偏移,在避免复杂的参数搜索过程的同时,保证了算法的鲁棒性能. 仿真结果表明,在相同初始条件下,改进算法较原算法对不规则分布数据的处理效率更高.

关键词: 一类支持向量机; 模糊隶属度; 核方法; 聚类
中图分类号: TP301 **文献标识码**: A

Improved kernel method for clustering based on fuzzy 1-SVM

XU Lei, ZHAO Guang-zhou
(College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China. Correspondent: XU Lei, E-mail: zhaogz@zju.edu.cn)

Abstract: By replacing the distance rule with fuzzy factors, an SVM featured fuzzy membership function of training points is introduced to work on the error coefficient. By punishing the weight of the remote points, the clustering center is prevented from being attracted by the abnormal data edge, thus the robustness against the remote points is improved without extra complex searching of parameters. Computer simulations show that, under the same initial conditions, the improved algorithm works more efficiently than the original one on irregularly distributed data.

Key words: 1-SVM; Fuzzy membership; Kernel method; Clustering

1 引 言

为了将支持向量机理论[1]应用于无指导学习, Tax 和 Duin[2], 以及 Scholkopf[3] 提出了一类支持向量机(1-SVM)算法,其核心思想是在高维特征空间计算包含输入数据映像的最小超球体. 该算法应用于层次聚类[4,5]、支持向量域描述[2]、类似 K 均值的迭代核聚类[6]等,均获得了良好的学习效果. 在核聚类算法中,文献[6]引入描述训练可行域的距离参数 ρ,在训练过程中拒绝所有与聚类中心的距离大于 ρ 的数据点. 这种方法虽然加强了算法对不规则分布数据的鲁棒性,但却带来了额外的参数选择计算,同时增加了迭代次数.

为了避免采用绝对的距离拒绝尺度,本文引进具有支持向量特性的模糊因子[6]来描述数据点属于聚类集的程度,并将作用于原算法中的误差系数形成模糊一类支持向量机(Fuzzy 1-SVM,或 1-FSVM). 模糊因子惩罚了边缘点对聚类中心的贡献权重,确保了聚类中心不会被个别边缘点所吸引. 仿真结果表明,模糊因子在功能上完全可以取代距离参数 ρ,从而在避免复杂的参数搜索过程的同时,保证了算法对不规则分布数据的鲁棒性.

2 基于 1-SVM 的核聚类算法

2.1 算法介绍

2.1.1 1-SVM 训练算法

一类支持向量机(1-SVM)是一种用支持向量描述数据点分布的方法,它试图用一个高维特征空间的超球体覆盖所有数据点在该特征空间的映像. 定义数据集 $D=\{x_i\}_{i=1}^n$, $x_i\in R^N$,设存在从 D 到某高维特征空间 χ 的非线性映射 ϕ,使得 $\phi(x_i)\in\chi$,寻找一个半径为 r 球心为 a 的超球体,使之尽可能覆盖 $\phi(x_i)$[1]. 即如下优化问题:

$$\min_{r,a,\xi_i} r^2 + C\sum_i \xi_i,$$

$$\text{s.t.} \quad \|\phi(x_i)-a\|^2 \le r^2+\xi_i,\ \xi_i\ge 0,\ \forall i.$$

其中: $\|\cdot\|$ 为欧氏距离; $\xi_i\ge 0$ 为松弛变量; C 为误差系数,用来调节误差与边界摆动之间的平衡[1]. 引入拉格朗日函数 $L_P(r,a,\xi_i)$,即最小化如下问题:

$$L_P(r,a,\xi_i) = r^2 - \sum_i \xi_i\beta_i + C\sum_i \xi_i -$$

收稿日期: 2007-05-31; **修回日期**: 2007-08-23.

作者简介: 徐磊(1981—),男,湖北汉川人,博士生,从事模式识别、支持向量机的研究; 赵光宙(1946—),男,浙江义乌人,教授,博士生导师,从事电气传动及其自动化、非线性系统控制等研究.

$$\sum_i (r^2 + \xi_i - \| \phi(x_i) - a \|^2)\alpha_i, \quad (1)$$

其中 $\alpha_i \geq 0$ 和 $\beta \geq 0$ 为拉格朗日乘子. 将 r, a 和 ξ_i 分别最小化,得

$$\sum_i \alpha_i = 1, \quad (2)$$

$$a = \sum_i \alpha_i \phi(x_i), \quad (3)$$

$$\alpha_i = C - \beta. \quad (4)$$

将式(2) ～ (4) 代入(1),可得 $L_P(r, a, \xi)$ 的 Wolfe 对偶型,则原问题可转化为如下标准二次规划问题:

$$\max_{\alpha_i} L_D = \sum_i \alpha_i K(x_i, x_i) - \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j),$$

$$\text{s.t. } 0 \leq \alpha_i \leq C, \quad \sum_i \alpha_i = 1. \quad (5)$$

且满足 KKT 条件 $(r^2 + \xi_i - \| \phi(x_i) - a \|^2)\alpha_i = 0$ 和 $\xi_i \beta_i = 0$,其中 K 为 Mercer 核函数,以替代空间 χ 上的内积[7],即 $K(x_i, x_j) = \phi(x_i)\phi(x_j)$.

训练完成后的数据点可分为 3 类: $\alpha_i = 0$ 对应的点为内点(个别点在球面上); $0 < \alpha_i < C$ 对应的点落在球面上,为内支持向量; $\alpha_i = C$ 对应的点落在球体外(个别点在球面上),为外支持向量. 由式(3)知,球心完全由支持向量所确定.

定义 $d^2(x) = \| \phi(x) - a \|^2$ 为任一数据点到球心的距离函数,由式(3) 和核函数的性质可得

$$d^2(x) = K(x, x) - 2\sum_i \alpha_i K(x_i, x) + \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j). \quad (6)$$

2.1.2　基于 1-SVM 的核聚类算法

文献[3] 将 1-SVM 成功应用于类似 K 均值的迭代式核聚类算法. 给定数据集 D 同上,考虑存在 L 个类的聚类问题($L \ll n$),设初始的聚类结果集为 $V = \{V_1, V_2, \cdots, V_L\}$,满足 $V_1 \cup V_2 \cup \cdots \cup V_L \subseteq D$ 和 $V_i \cap V_j = \varnothing$,其中 $i, j \in 1, 2, \cdots, L$ 且 $i \neq j$. 将 $V_1, V_2, \cdots, V_L$ 分别进行 1-SVM 训练,得到 L 个超球体,则聚类中心集为 $W = \{w_i\}_{i=1}^{L}$,其中 w_i 为第 i 个类的超球体的球心. 新的聚类结果集按数据点到聚类中心的距离重新划分,即

$$V_l = \{\phi(x_i) \in \chi \mid l = \arg \min_{j=1,2,\cdots,L} \| \phi(x_i) - \phi(w_j) \|\},$$

$$l = 1, 2, \cdots, L, \ i = 1, 2, \cdots, n. \quad (7)$$

其中 $\| \phi(x_i) - \phi(w_j) \|$ 由式(6) 计算得出. 至此,便完成了一步迭代. 综上,基于 1-SVM 的迭代式核聚类算法的完整步骤如下:

1) 任选数据集 D 的 L 个子集 $V_1, V_2, \cdots, V_L$ 作为初始聚类结果集,满足 $V_1 \cup V_2 \cup \cdots \cup V_L \subseteq D$ 和 $V_i \cup V_j = \varnothing$,其中 $i, j \in 1, 2, \cdots, L$ 且 $i \neq j$.

2) 对每个 V_l 使用 1-SVM 训练. 聚类中心集 W 更新为对应超球体的球心.

3) 根据式(6) 和(7) 更新 V.

4) 如果 3) 中 V 已无变化,则算法结束,最终聚类结果集为 V,聚类中心集为 W;否则转到 2).

2.2　问题分析

1-SVM 算法选取高斯或拉普拉斯核函数比较合适[6],本文采用高斯核,即

$$K(x_i, x_j) = \exp(-q \| x_i - x_j \|^2),$$

其中 $q \in R$. 算法的训练结果直接受参数 C 和 q 的影响[4,5]. q 值可由数据点的先验信息经验性地获得,原则上应尽量避免 $q \| x_i - x_j \|^2$ 太大或太小;误差参数 C 约束了 α_i 的取值上界,若 $C \geq 1$,则不产生外支持向量. 为了简化参数选择,文献[3] 令 $C \equiv 1$,仅调整参数 q. 为避免被偏远的边缘点所干扰,文献[3] 为每个聚类目标引进参数 $\rho_l (l = 1, 2, \cdots, L)$ 作为距离拒绝尺度. 在更新聚类结果时,凡是与第 l 个球心的距离大于 ρ_l 的点,将被排除在聚类结果集 V_l 之外,即

$$V_l = \{\phi(x_i) \in \chi \mid l = \arg \min_{j=1,2,\cdots,L} \| \phi(x_i) - \phi(w_j) \|, \ \| \phi(x_i) - \phi(w_l) \| \leq \rho\}.$$

该方法虽然在一定程度上加强了算法对不规则分布数据的鲁棒性,但仍存在以下问题:1) 确定参数 ρ 需要采用模型选择技术[9],此过程基本等同于计算最终聚类结果集的半径,以此作为先验信息过于强大;2) 对于不同的初始点,参数 ρ 要相应作出调整,否则可能导致不同的聚类结果;3) 由于每次训练只有部分满足距离尺度的数据点归于聚类结果集,导致算法收敛所需的迭代次数增加.

3　改进算法

3.1　模糊一类支持向量机

为削弱边缘点对聚类结果的影响,除了设定一个绝对的距离尺度外,还可考虑模糊关系来描述边缘点被排除在聚类集之外的程度. 由式(3) 知,在球心的组成部分中,外支持向量分得了恒定的最大权重 C,内支持向量获得的权重小于 C,这将导致在迭代过程中球心被边缘点所吸引. 若定义模糊因子以描述训练点隶属于聚类集的程度,将其作用于 C,则球心构成的权重可随着数据点的偏远程度而变化;同时,C 作为误差系数的意义也更加清晰:对边缘点投入较少的关注,而将误差惩罚的重心转移到更可能属于该聚类集的数据点上.

在 1-SVM 算法的基础上,借鉴文献[10] 提出的模糊 SVM 模型,引进应用于迭代式核聚类算法的模糊一类支持向量机(1-FSVM). 定义点 x_i 属于原

超球体的模糊隶属度为 $s_i(0 < s_i < 1)$[6],则在下一轮迭代训练中新的最优化问题为

$$\min_{r,a,\xi_i} r^2 + C\sum_i s_i\xi_i,$$
$$\text{s.t.} \quad \|\phi(x_i) - a\|^2 \le r^2 + \xi_i,$$
$$\xi_i \ge 0, \ \forall i. \tag{8}$$

参照式(1),拉格朗日函数为

$$L_{SP}(r,a,\xi_i) = r^2 - \sum_i \xi_i\beta_i + C\sum_i \xi_i s_i - \sum_i (r^2 + \xi_i - \|\phi(x_i) - a\|^2)\alpha_i. \tag{9}$$

极小化 L_{SP},仍得到式(2)和(3),原式(4)则变为

$$\alpha_i = Cs_i - \beta_i. \tag{10}$$

代入式(8),可消去 s_i,得到二次规划模型

$$\max_{\alpha_i} L_{SD} = \sum_i \alpha_i K(x_i, x_i) - \sum_{i,j}\alpha_i\alpha_j K(x_i, x_j),$$
$$\text{s.t.} \ 0 \le \alpha_i \le Cs_i, \ \sum_i \alpha_i = 1. \tag{11}$$

该算法本质上是针对每个数据点可调整误差参数 C 的 1-SVM. 不同于文献[10]仅聚类一次即结束,本文将模糊算法的特性充分应用于迭代过程中,由每次迭代结束后的分类关系生成模糊隶属度,并参与到下一次的迭代计算中. 只要设定了合理的模糊隶属度法则,误差参数便可根据数据点对聚类中心的分布特性作出自适应的调整,从而逐步地削弱边缘点对聚类结果的影响. 因此,完整的聚类算法的另一个关键步骤在于构造合适的模糊隶属度函数.

3.2 模糊隶属度函数

模糊隶属度函数 s_i 的一种流行算法[6]是选取数据点和球心的距离比例. 定义类别半径 $r' = \max\|\phi(x_i) - a\|$,则 s_i 可表示为 $s_i = 1 - \|\phi(x_i) - a\|/r'$. 此定义存在两个问题:分类的可信度不高,未充分利用最小超球体的训练成果;未考虑数据点之间的紧密性质,如图 1 所示.

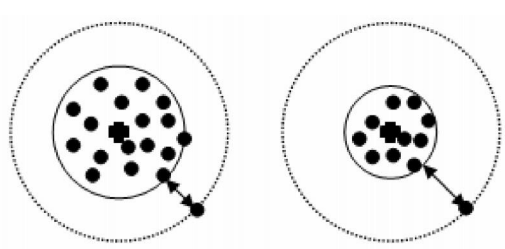


图 1 相同中心距不同紧密度的数据点

在已知 1-SVM 训练结果的基础上定义模糊隶属度函数,可利用已最优化的最小超球体作为隶属程度的参照. 参考式(6),令 $d(x_i) = \|\phi(x_i) - a\|$,则模糊隶属度函数可定义如下[11]:

$$s_i = \begin{cases} 0.5 \times \dfrac{1 - d(x_i)/r}{1 + \lambda_1 d(x_i)/r} + 0.5, & d(x_i) \le r; \\ 0.5 \times \dfrac{1}{1 + \lambda_2(d(x_i) - r)}, & d(x_i) > r. \end{cases} \tag{12}$$

为使模糊隶属度函数在 $d(x_i) = r$ 时可微,参数 λ_1 和 λ_2 满足 $r(1 + \lambda_1)\lambda_2 = 1$. 本文 λ_1 取值为 1. 可见,内点和内支持向量属于该分类的模糊程度均大于 0.5,而外支持向量则倾向于远离该分类.

另外,原算法在步骤 3)中需利用式(6)和(7)更新聚类结果集 V,该步骤必须遍历全体数据点并计算相应的 $d(x_i)$. 因此,在单步迭代内,模糊隶属度的计算并不会增加算法的渐近时间复杂度.

3.3 算法步骤

综上,基于 1-FSVM 的迭代式核聚类算法步骤如下:

1) 任选数据集 D 的 L 个子集 $V_1, V_2, \cdots, V_L$ 作为初始聚类结果集,满足 $V_1 \cup V_2 \cup \cdots \cup V_L \subseteq D$ 和 $V_i \cap V_j = \varnothing$,其中 $i, j \in 1,2,\cdots,L$ 且 $i \ne j$. 令所有初始点对应的 $s_i = 1$.

2) 对每个 V_l 使用 1-FSVM 训练. 聚类中心集 W 更新为对应超球体的球心.

3) 根据式(6)和(7)更新 V.

4) 如果在 3)中 V 已无变化,则算法结束,最终聚类结果集为 V,聚类中心集为 W;否则,根据式(12)对每个 V_l 更新 s_i,转到 2).

4 仿真结果和讨论

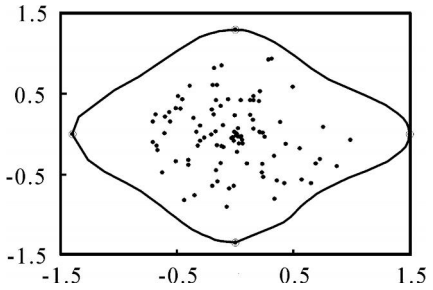
4.1 单步聚类

为检验两种算法对边缘点的作用,构造如下算例:第 1 步训练对象为单位圆内随机选取的 100 个点,假设训练完成后更新的数据集增加了在单位圆外的 4 个边缘点:(1.5,0),(-1.4,0),(0,1.3),(0,-1.35),则两种算法第 2 步的训练效果如图 2 所示(为便于比较,训练参数均为 $C = 1$ 和 $q = 0.6$),其中加圆圈的点为支持向量.

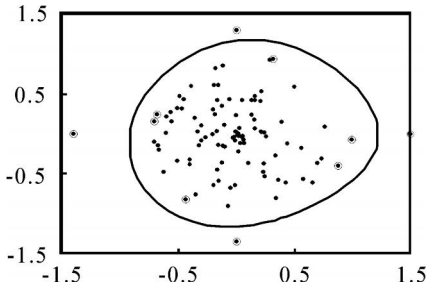
可见,1-SVM 的训练结果完全被边缘点所支配(仅有的 4 个支持向量全是新增的边缘点),而上一步聚类的单位圆信息几乎全部丢失. 1-FSVM 在一定程度上削弱了边缘点的干扰,全部 4 个边缘点被成功地划分为外支持向量,且内支持向量基本上保留了原单位圆的信息.

4.2 迭代聚类

分别将基于 1-FSVM 的核聚类算法、基于 1-SVM 的核聚类算法和经典 K 均值算法应用于对 IRIS 数据集的迭代聚类. 为使结果可视化,IRIS 集经主成分分析降为二维. 其中,1-SVM 未引进参数



(a)　1-SVM 的训练效果

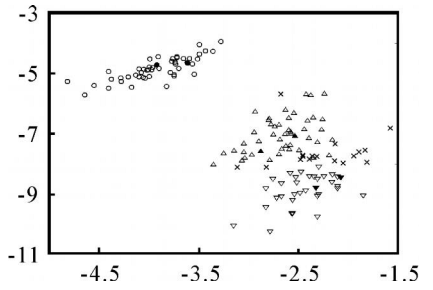


(b)　1-FSVM 的训练效果

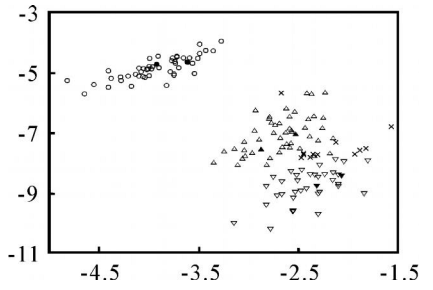
图 2　单步聚类效果

ρ(一个足够准确的先验参数 ρ会产生良好的结果),3 种算法均采用相同的小容量初始集(较大容量的初始集含有太多先验信息,可能影响聚类结果),两种核聚类的 SVM 参数均为 $C = 1$ 和 $q = 0.6$. 算法最终的聚类结果如图 3 所示,数据点分为圆、上三角、下三角 3 类,其中初始集用实心点表示,误分类点用实心叉表示. 3 种算法的准确率分别为:图 3(a) 准确率为 88 %,图 3(b) 准确率为 91.3 %,图 3(c) 准确率为 96.7 %. 可见,带模糊因子的核聚类方法比未引入参数 ρ的核聚类方法能获得更高的准确率.

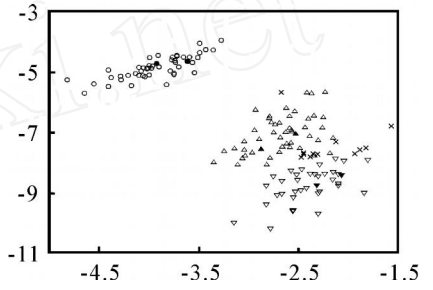
表 1 给出了基于无距离参数的 1-SVM 聚类算法(简称算法 1)、基于 1-FSVM 的聚类算法(简称算法 2) 以及基于有距离参数的 1-SVM 聚类算法(简称算法 3) 在收敛性质和准确率上的比较. 测试数据集仍采用与图 3 相同的 IRIS 数据集;4 组初始集均为随机获取的小容量集合,每个聚类目标仅包含 2 个初始点;距离参数取值为 $\rho_1 = 1.50$, $\rho_2 = 0.91$, $\rho_3 = 1.69$. 在准确率方面,算法 2 和算法 3 差别不大,且均优于算法 1,表明距离参数和模糊因子均能改善聚类准确率. 在时间消耗上,算法2略优于算



(a)　K 均值聚类



(b)　基于 1-SVM 的迭代核聚类



(c)　基于 1-FSVM 的迭代核聚类

图 3　聚类结果

法 3,其原因在于:在单步迭代的渐近时间复杂度相同的前提下,算法 2 的迭代次数略少于算法 3(见2.2 节问题分析). 考虑到选择参数 ρ需要额外的计算开销(不论采用贝叶斯方法或交叉验证法[9],模型选择法均需在给定的训练集内反复计算各种参数条件下的泛化误差),因而本文提出的算法 2 比算法 3 更容易实现.

可见,模糊因子在功能上类似于距离参数 ρ,能使算法以牺牲一定的收敛速度来换取更高的聚类准确率. 但是模糊算法并不需要通过模型选择法反复搜索合适的距离参数,因此在不失鲁棒性的条件下该算法具有更高的效率和更强的实用性.

表 1　不同初始条件下算法的迭代次数和准确率

初始集	无距离参数 1-SVM			1-FSVM			有距离参数 1-SVM		
	迭代次数	计算时间 s	准确率 %	迭代次数	计算时间 s	准确率 %	迭代次数	计算时间 s	准确率 %
1	2	1.08	91.3	5	3.21	96.7	6	3.52	95.3
2	3	1.40	84.7	3	2.31	89.3	5	3.02	90.7
3	3	1.73	94	4	2.73	94	6	3.41	94
4	2	0.89	89.3	6	4.37	97.3	6	3.32	95.3

5 结 论

本文提出一种基于 1-FSVM 的迭代式核聚类算法,将具有支持向量特性的模糊隶属度函数作用于误差系数,惩罚了边缘点对聚类的影响,保证了算法的鲁棒性能.仿真实验表明,1-FSVM 在单步聚类时可显著地惩罚边缘点,并保证原聚类集合的性质;在迭代聚类时,模糊因子可达到与距离参数类似的消除边缘点影响的效果,通过增加收敛次数获得更高的分类准确率,并且避免了复杂的参数搜索过程.因此,改进的算法较原算法对不规则分布数据的处理效率更高,更具实用性.

进一步的工作是研究参数和收敛速度的相对关系,并致力于该算法的广泛外推.

参考文献(References)

[1] Tax D M J, Duin R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11-13): 1191-1199.

[2] Scholkopf B, Williamson R C, Smola A J, et al. Support vector method for novelty detection [J]. Advances in Neural Information Processing Systems, 1999, 12: 526-532.

[3] Camastra F, Verri A. A novel kernel method for clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(5): 801-805.

[4] Ben-Hur A, Horn D, Siegelmann H T, et al. Support vector clustering [J]. Machine Learning Research, 2001, 2: 125-137.

[5] Ben-Hur A, Horn D, Siegelmann H T, et al. A support vector method for hierarchical clustering[J]. Advances in Neural Information Processing Systems, 2001, 13: 367-373.

[6] 李炜, 黄心汉, 陈曦. 一种基于模式类特征空间统计分布的离散模糊隶属度函数模型[J]. 信号处理, 2004, 20(2): 170-173.
(Li W, Huang X H, Chen X. A model of discrete fuzzy membership function based on statistical distribution of features of pattern[J]. Signal Processing, 2004, 20(2): 170-173.

[7] Saitoh S. Theory of reproducing kernels and its applications [M]. London: Longman Scientific & Technical, 1988.

[8] Vapnik V. Statistical learning theory[M]. New York: Wiley, 1998.

[9] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning[M]. New York: Springer-Verlag, 2001.

[10] Wei L L, Long W J, Zhang W X. Fuzzy data domain description using support vector machines[C]. The 2nd Int Conf on Machine Learning and Cybernetics. Xi'an, 2003: 3802-3805.

[11] Chiang J H, Hao P Y. A new kernel-based fuzzy clustering approach: Support vector clustering with cell growing[J]. IEEE Trans on Fuzzy Systems, 2003, 11(4): 518-527.

下 期 要 目