

基于 FCM 隶属度的支持向量机

齐丙娟¹, 丁世飞^{1,2}

(1 中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116;

2 中国科学院 计算技术研究所 智能信息处理重点实验室, 北京 100190)

摘 要: 传统 SVM 在训练大规模数据集时, 训练速度慢, 时间消耗代价大. 针对此问题, 提出利用 FCM 算法对训练样本集进行预处理, 依据样本隶属度提取出所有可能的支持向量进行 SVM 训练. 利用原始数据集对算法进行验证, 此算法在保证 SVM 分类精度的同时, 大大提高了训练速度, 算法具有可行性.

关键词: 统计学习理论; 支持向量机; 模糊 C 均值聚类; 隶属度; 分类精度; 训练速度

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2011)10-0048-04

Support Vector Machine Based on Memberships of FCM

QI Bing-juan¹, DING Shi-fei^{1, 2}

(1 School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China;

2 Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,

Chinese Academy of Science, Beijing 100190, China)

Abstract: As far as we know, circuit design is a process of cumbersome, complex and long cycle. This paper provides a new modeling approach based on evolutionary for circuit networks, which converts the process of network modeling into the optimization process of the determination of circuit components and component values. The new method is proposed for hybrid identification of circuit components and component values by performing global optimal search in the complex solution space and achieving the evolution of circuit networks through the way of genetic programming.

Key words: statistical learning theory; support vector machine; fuzzy C-means clustering; memberships; classification accuracy; training speed

1 引言

统计学习理论^[1-2]是 Vapnik 等人提出的一种小样本统计理论, 着重研究在小样本情况下的统计规律及学习方法性质. 支持向量机^[3-4] (Support Vector Machine, SVM) 是在统计学习理论 VC 维理论和结构风险最小化原理的基础上发展起来机器学习方法. 由于它能够较好地处理小样本、非线性、维数灾难、过学习和局部极小等问题, 再加上它具有坚实的理论基础, 简单明了的数学模型, 因此在模式识

别、回归分析、函数估计、时间序列预测等领域都得到了长足的发展.

尽管 SVM^[5] 近些年的发展迅速, 在 SVM 训练算法^[6-7] 的改进和 SVM 的应用^[8-9] 等方面都取得了令人瞩目的成果, 但是在解决实际问题时, SVM 的训练速度极大地受到训练集规模的影响. 在支持向量机对数据样本集进行学习之前, 适当进行数据预处理, 可以提高训练速度和分类精度. 一般处理方法主要有以下几类: 对输入样本进行特征选择, 即通过减少输入样本的维数, 提高分类器的分类能力, 如文

收稿日期: 2011-04-20; 修回日期: 2011-06-28

基金项目: 国家自然科学基金(60975039); 江苏省基础研究计划(自然科学基金)(BK2009093); 中国科学院智能信息处理重点实验室开放基金项目(IIP2010-1)

献[10]中提出的 RS-SVM,文献[11]中提出的基于遗传算法的特征选择;通过将原本大学习样本集分割成多个小的样本集合,以减少训练复杂度,如文献[12]中提出的分布式 SVM. 本文采用模糊 C 均值聚类算法(FCM)对训练数据集进行预处理,根据隶属度的范围提取出所有可能是支持向量(Support Vector,SV)的样本作为 SVM 的训练样本集,通过实验验证此方法在保证 SVM 分类精度的同时,缩减了训练样本集的规模,提高了训练速度.

2 支持向量机

支持向量机(SVM)是建立在统计学习理论基础上的—种数据挖掘方法,其机理是寻找一个满足分类要求的最优分类超平面,使得该超平面在保证分类精度的同时,能够使超平面两侧的空白区域最大化.

以两类数据分类为例,给定训练样本集 $(x_i, y_i), i = 1, 2, \dots, l, x \in R^n, y \in \{\pm 1\}$, 超平面记作 $(w \cdot x) + b = 0$, 为使所有样本正确分类并且具备最大分类间隔,构造最优超平面的问题即在约束条件下求:

$$\min \Phi(x) = \frac{1}{2} \|w\|^2 = \frac{1}{2} (w \cdot w) \quad (1)$$

为了解决这个最优化问题,引入 Lagrange 函数:

$$L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^l a_i (y_i ((w \cdot x_i) + b) - 1) \quad (2)$$

式中, $a_i > 0$ 为 Lagrange 乘数.

相应的对偶问题为

$$\max Q(a) = \sum_{j=1}^l a_j - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l a_i a_j y_i y_j (x_i \cdot x_j) \quad (3)$$

$$\text{s. t. } \sum_{j=1}^l a_j y_j = 0, j = 1, 2, \dots, l$$

$$a_j \geq 0, j = 1, 2, \dots, l$$

最优分类函数为

$$\begin{aligned} f(x) &= \text{sgn}\{(w^* \cdot x) + b^*\} \\ &= \text{sgn}\{(\sum_{j=1}^l a_j^* y_j (x_j \cdot x_i)) + b^*\}, x \in R^n \end{aligned} \quad (4)$$

对于线性不可分情况,SVM 的主要思想是将输入向量映射到高维特征向量空间中,并在该特征空间中构造最优分类面.

将 x 作从输入空间 R^n 到特征空间 H 的变换:

$$x \rightarrow \Phi(x) = (\Phi_1(x), \Phi_2(x), \dots, \Phi_l(x))^T \quad (5)$$

得到最优分类函数为

$$\begin{aligned} f(x) &= \text{sgn}(w \cdot \Phi(x) + b) \\ &= \text{sgn}(\sum_{i=1}^l a_i y_i \Phi(x_i) \cdot \Phi(x) + b) \end{aligned} \quad (6)$$

3 模糊 C 均值聚类

模糊 C 均值聚类^[14-15](FCM) 算法是一种基于划分的聚类算法,它的思想就是使得被划分到同一簇的对象之间相似度最大,而不同簇之间的相似度最小.

FCM 把 n 个向量 $x_i (i = 1, 2, \dots, n)$ 分为 c 个模糊组,并求每组的聚类中心,使得非相似性指标的价值函数达到最小. FCM 用模糊划分使每个给定数据点用值为 $0 \sim 1$ 的隶属度来确定其属于各个组的程度. 与引入模糊划分相适应,隶属矩阵 U 允许有取值为 $0 \sim 1$ 间的元素,并且一个数据集的隶属度的和总等于 1:

$$\sum_{i=1}^c u_{ij} = 1, \forall j = 1, \dots, n \quad (7)$$

那么,FCM 的价值函数(或目标函数):

$$J(U, c_1, \dots, c_c) = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2 \quad (8)$$

这里 u_{ij} 介于 $0 \sim 1$, c_i 为模糊组 I 的聚类中心, $d_{ij} = \|c_i - x_j\|$ 为第 I 个聚类中心与第 j 个数据点之间的欧几里德距离, $m \in [1, \infty)$ 是一个加权指数.

构造如下新的目标函数,可求得使式(8)达到最小值的必要条件:

$$\begin{aligned} \bar{J}(U, c_1, \dots, c_c, \lambda_1, \dots, \lambda_n) \\ = J(U, c_1, \dots, c_c) + \sum_{j=1}^n \lambda_j (\sum_{i=1}^c u_{ij} - 1) \quad (9) \\ = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2 + \sum_{j=1}^n \lambda_j (\sum_{i=1}^c u_{ij} - 1) \end{aligned}$$

这里 $\lambda_j (j = 1, 2, \dots, n)$ 是式(9)的 n 个约束式的拉格朗日乘子. 对所有输入参量求导,使式(10)达到最小的必要条件为

$$c_i = \frac{\sum_{j=1}^n u_{ij}^m x_j}{\sum_{j=1}^n u_{ij}^m} \quad (10)$$

和

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \quad (11)$$

由上述两个必要条件,FCM 是一个简单的迭代过程.

4 基于 FCM 的 SVM

对一些大规模数据集,直接采用 SVM 进行训练会有很大的时间消耗,因此考虑对原始数据集进行预处理^[16-18],通过提取所有可能的支持向量来缩减样本规模,以达到提高 SVM 训练速度的目的. FCM 依据样本隶属度进行聚类,每一类别的样本隶属度取值不等,但均取值为 0~1. 样本隶属度越接近于 1,则该样本为支持向量的概率越小,样本隶属度越接近于 0,则该样本为支持向量的概率越大. 依据这个结论,本文算法中对隶属度的取值有 0. 95、0. 9、0. 85、0. 8,对这四个给定隶属度进行验证算法的可行性.

基于 FCM 隶属度的 SVM 算法描述:

步骤 1:用值为 0~1 的随机数初始化隶属矩阵 U ,使其满足式(9) 中的约束条件.

步骤 2:用式(10) 计算 c 个聚类中心 $c_i(i=1,2,\cdots,c)$.

步骤 3:根据式(8) 计算价值函数. 如果它小于某个确定的阈值,或它相对上次价值函数值的改变量小于某个阈值,转步骤 5.

步骤 4:用式(11) 计算新的 U 矩阵. 返回步骤 2.

步骤 5:聚类算法结束,根据样本隶属度,分别输出隶属度小于 0. 95、0. 9、0. 85、0. 8 的样本数据.

步骤 6:对输出样本数据分别进行 SVM 训练,用原始数据集检验其分类正确率.

5 实验结果分析

本文采用 iris 数据集、wine 数据集、ASD-12-2 数据集和 ASD-14-2 数据集进行测试算法的可行性(表 1).

表 1 数据集概况

数据集	样本个数	类别个数	属性个数
Iris	150	3	4
Wine	178	3	13
ASD-12-2	535	12	2
ASD-14-2	685	14	2

由表 2 可见,采用 FCM 对数据集进行处理之后,训练样本集的样本数大大减少,达到缩减训练样本集的目的,提高了 SVM 的训练速度,同时 SVM 的分类精度也没有很大降低. 尤其,在隶属度取值

0. 95 时,四个实验数据集都能保持原有的分类精度,在隶属度取值 0. 9、0. 85、0. 8 时,四个实验数据集的分类精度也没有很大降低. 因此,该算法具有可行性.

表 2(1) FCM+SVM 训练结果

数据集	SVM		隶属度 $u<0. 95$	
	SV 数/训练样本数	分类正确率/%	SV 数/训练样本数	分类正确率/%
Iris	44/150	98. 67	43/116	98. 67
Wine	178/178	100	178/178	100
ASD-12-2	91/535	100	89/424	10
ASD-14-2	110/685	100	109/580	100

表 2(2) FCM+SVM 训练结果

数据集	SVM		隶属度 $u<0. 9$	
	SV 数/训练样本数	分类正确率/%	SV 数/训练样本数	分类正确率/%
Iris	44/150	98. 67	42/88	98. 67
Wine	178/178	100	178/178	100
ASD-12-2	91/535	100	85/339	100
ASD-14-2	110/685	100	105/482	100

表 2(3) FCM+SVM 训练结果

数据集	SVM		隶属度 $u<0. 85$	
	SV 数/训练样本数	分类正确率/%	SV 数/训练样本数	分类正确率/%
Iris	44/150	98. 67	41/65	99. 33
Wine	178/178	100	176/176	98. 88
ASD-12-2	91/535	100	74/272	92. 52
ASD-14-2	110/685	100	93/381	94. 16

表 2(4) FCM+SVM 训练结果

数据集	SVM		隶属度 $u<0. 8$	
	SV 数/训练样本数	分类正确率/%	SV 数/训练样本数	分类正确率/%
Iris	44/150	98. 67	35/50	98. 67
Wine	178/178	100	169/169	95. 51
ASD-12-2	91/535	100	67/229	92. 34
ASD-14-2	110/685	100	85/310	94. 01

6 结束语

随着 Internet 及各种信息系统的迅猛发展,产

生了海量、高维、分布式、动态等复杂数据,传统支持向量机在处理大规模高维复杂数据时存在很多弊端,例如训练速度过慢,时间消耗代价过大,分类精度不足等.针对这些问题,国内外很多学者提出了许多解决方案,在其基础之上,本文提出基于模糊C均值聚类隶属度的支持向量机,通过利用FCM算法对训练样本集进行预处理,提取出所有可能的支持向量进行SVM训练,并使用原始数据集对训练结果进行验证.经实验证明,本文提出的算法在保证SVM分类精度的同时,大大提高了训练速度.

参考文献:

- [1] Vapnik V N. 统计学习理论的本质[M]. 张学工,译. 北京:清华大学出版社,2000.
- [2] Vapnik V N. 统计学习理论[M]. 许建华,张学工,译. 北京:电子工业出版社,2004.
- [3] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):32-41.
- [4] Cristianini N, Shawe-Taylor J. 支持向量机导论[M]. 李国正,王猛,译. 北京:电子工业出版社,2004.
- [5] 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011,40(1):2-10.
- [6] 王书舟,伞冶. 支持向量机的训练算法综述[J]. 智能系统学报,2008,3(6):467-475.
- [7] 文益民,王耀南,吕宝粮,等. 支持向量机处理大规模问题算法综述[J]. 计算机科学,2009,36(7):20-25.
- [8] 祁亨年. 支持向量机及其应用研究综述[J]. 计算机工程,2004,30(10):6-9.
- [9] 刘晓亮,丁世飞. SVM用于文本分类的适用性[J]. 计算机工程与科学,2010,32(6):106-108.
- [10] 张国云,章兢. 基于粗集理论和支持向量机的动态预测新方法及应用[J]. 自动化仪表,2005,26(2):17-20.
- [11] 赵倩. 基于SVM和遗传算法的皮肤显微图像特征选择[J]. 模式识别与人工智能,2005,18(4):502-505.
- [12] 桂卫华,李勇刚. 基于改进聚类算法的分布式SVM及其应用[J]. 控制与决策,2004,19(8):852-856.
- [13] 袁亚湘. 最优化理论与方法[M]. 北京:科学出版社,1997.
- [14] 高新波. 模糊聚类分析及其应用[M]. 西安:西安电子科技大学出版社,2004.
- [15] 刘笛,朱学峰,苏彩红. 一种新型的模糊c均值类初始化方法[J]. 计算机仿真,2004,21(11):148-151.
- [16] 李园. 基于模糊聚类的支持向量机的分类算法研究[D]. 哈尔滨:哈尔滨工业大学,2005.
- [17] 张良春,夏利民,石华玮. 基于模糊聚类支持向量机的高速公路事件检测[J]. 计算机工程与应用,2007,43(7):206-209.
- [18] 曹晓莉,江朝元,甘思源. 基于聚类支持向量机的船用污水处理装置故障诊断[J]. 计算机应用,2008,28(10):2648-2651.

作者简介:

齐丙娟 女,(1987-),硕士研究生. 研究方向为支持向量机.

丁世飞 男,(1963-),博士后,教授,博士生导师. 研究方向为模式识别与人工智能、机器学习与数据挖掘等.

(上接第47页)

的特点.能提高复杂垂直边缘背景干扰下候选车牌提取的准确性,从而进一步提高整体车牌定位的检测率.实验结果表明算法对图像背景、车牌大小和位置的限制较少.该方法简单、高效、具有较强的通用性和较高的实用价值.

参考文献:

- [1] 苑玮琦,张亮. 基于分块思想的汽车牌照定位算法研究[J]. 自动化学报,2007,33(7):768-770.
- [2] 吴舟舟,李树广. 基于分级边缘间距的实时车牌检测[J]. 中国图象图形学报,2007,12(2):315-321.
- [3] 片兆宇,孟祥萍,张红. 多阶段车牌定位算法[J]. 计算机科学,2009,36(8):297-299.
- [4] 黄豪杰,李榕,常鸿森. 基于自适应灰度形态学滤波的车牌图像分割[J]. 微电子学与计算机,2007,24(4):95-98.
- [5] Qian Cui, Suihe Ge, Shujian Shi, et al. License plate location based on direction and edge information[C]// Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications. China: Wuhan, 2009:300-303.
- [6] Sobel I E. Camera models and machine perception [D]. California: Standford University,1970.
- [7] Otsu N. A threshold selection method from gray-level histograms[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979,9(1):62-66.

作者简介:

葛建芳 女,(1969-),副教授. 研究方向为图像处理与模式识别.

朱玲玲 女,(1977-),讲师. 研究方向为图像处理与模式识别.