DOI: 10.13382/j.jemi.2015.04.018

基于粒子群算法的决策树 SVM 多分类方法研究*

王道明 鲁昌华 蒋薇薇 肖明霞 李必然
(合肥工业大学计算机与信息学院 合肥 230009)

摘 要: 针对 SVM 多分类问题提出了一种基于粒子群算法的最优决策树 SVM 生成算法,以解决传统支持向量机多分类方法存在的不可分区域和误差积累现象。该方法利用自变异的 PSO 聚类算法在每一决策节点自动寻找最优或近优分类决策,将数据集划分为两类,直至叶子节点为止,最终根据最优决策树构建 SVM 多分类结构,训练各个节点 SVM 分类器。将该算法应用于图像人群密度分类问题,仿真实验表明,分类精度和分类时间得到明显改善,是一种有效地的多分类算法。

关键词: 支持向量机; 粒子群算法; 决策树; 多分类
中图分类号: TN06 **文献标识码:** A **国家标准学科分类代码:** 520.2010

Study on PSO-based decision-tree SVM multi-class classification method

Wang Daoming Lu Changhua Jiang Weiwei Xiao Mingxia Li Biran
(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract: This paper proposes a PSO-based decision tree SVM multi-class classification algorithm to resolve unclassifiable region and error accumulation phenomenon existing in traditional support vector machine multi-class classification method. PSO algorithm is used to cluster the dataset into two patterns on each node, search optimal or suboptimal decision-tree automatically, and then construct the SVM classifier with the optimal decision tree. The algorithm is applied to the crowd density image classification problem. The simulation results show that classification accuracy and time are improved obviously, and it is an effective multi-class classification algorithm.

Keywords: support vector machine; particle swarm optimization; decision tree; multi-class classification

1 引 言

SVM 多分类问题是当前机器学习领域研究的热点之一,目前 SVM 多分类方法可分为两类: 1) 直接在目标函数上进行修改,将多个分类面的参数求解合并到一个最优化问题中,通过求解该最优化问题"一次性"实现多类分类; 2) 通过组合多个二分类器来实现多分类器的构造。第一类方法计算复杂度较高,实现较困难,只适合用于小型问题中。在实际中第二类方法更常用,其主要方法有 one-against-one (1-a-1),one-against-rest (1-a-r),决策树 SVM (DT-SVM),决策导向无环图 SVM (DAG-SVM) 以及纠错编码 SVM (ECOC-SVM) 等。

1-a-1 和 1-a-r 2 种方法的优点是较为简单,但是均存在不可分区域,其泛化误差无界,并且所需训练的基本 SVM 较多,影响训练和分类速度。DAG-SVM 以及 ECOC-SVM 2 种方法相对较为复杂,有文献 [4] 通过一系列实验比较指出简单的多分类方法 1-a-1, 1-a-r 比复杂的 DAG-SVM, ECOC-SVM 更适用于实际应用。决策树 SVM 方法需要构造的分类器少且不存在不可分区域,分类时不需经历全部分类器,是解决 SVM 多分类问题的优秀方法。针对 DT-SVM, 文献 [5] 使用超球体 (或超长方体) 作为样本分布范围度量最小包含某一类样本; 文献 [6] 把类 i 和类 j 中 2 个最近样本向量之间的欧氏距离作为类 i 和类 j 之间的距离; 文献 [7] 利用遗传算法自适应生成

收稿日期: 2014-07 Received Date: 2014-07
* 基金项目: 合肥工业大学承担安徽省科技强警 (1301b042014) 项目

SVM 决策树。

本文提出一种基于粒子群算法的决策树 SVM 多分类方法，应用于图像人群密度分类问题，实验结果表明，新方法具有较高的分类精度和较短的分类时间。

2 支持向量机

2.1 支持向量机算法

支持向量机[7]是基于统计学习理论的一种机器学习方法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的，基本思想是寻找最优分类面使正负类之间的分类间隔(margin)最大。

设训练样本为(x_i, y_i)，$i = 1,2,\cdots,l$，$x \in R^n$，$y \in \{+1, -1\}$，l为样本数，n为输入维数。当线性可分时，最优分类超平面为:

$$wx + b = 0 \tag{1}$$

此时分类间隔为$2/\|w\|$，显然当$\|w\|$值最小时，分类间隔最大。可以把问题描述为求解下述约束性优化问题:

$$\min \|w\|^2/2 \tag{2}$$
$$\text{s.t. } y_i(w \cdot x_i + b) - 1 \geqslant 0, \quad i = 1,2,\cdots,l$$

当训练样本集线性不可分时，需引入非负松弛变量ξ_i，$i = 1,2,\cdots,l$求解最优分类面问题为:

$$\min \|w\|^2/2 + C\sum_{i=1}^{l}\xi_i \tag{3}$$
$$\text{s.t. } y_i(w \cdot x_i + b) \geqslant 1 - \xi_i, \xi_i \geqslant 0, \quad i = 1,2,\cdots,l$$

式中: C为惩罚参数，C越大表示对错误分类的惩罚越大。通过 Lagrange 乘子法求解上述优化问题，可得最优决策函数为:

$$f(x) = \text{sgn}\left[\sum_{i=1}^{l} y_i a_i (x \cdot x_i) + b\right] \tag{4}$$

式中: a为 Lagrange 系数。在对输入测试样本x进行测试时，由式(4)确定x的所属类别。根据 K-T 条件，上述优化问题的解必须满足。

$$a_i(y_i(w \cdot x + b) - 1) = 0 \tag{5}$$

因此，对于多数样本a_i将为零，只有支持向量的a_i不为零，它们通常在全体样本中所占的比例很少。这样，仅需要少量支持向量即可完成正确的样本分类。

非线性分类问题时，SVM 通过核函数$K(x_i \cdot x_j)$将样本x映射到某个高维空间H，然后在H中对原始问题进行线性划分。根据 Mercer 条件，此时相应的最优决策函数变为:

$$f(x) = \text{sgn}\left[\sum_{i=1}^{l} y_i a_i K(x \cdot x_i) + b\right] \tag{6}$$

2.2 决策树支持向量机

决策树支持向量机[8]首先将所有类别分为 2 个子类，每个子类在下一层次继续划分为 2 个次子类，如此循环，直至生成叶子节点，即只包含一个单独类的节点，最终形成决策树，每个决策点使用 SVM 进行分类。

决策树 SVM 与 1-a-1 方法及 1-a-r 方法相比具有以下优点:

1) 决策树 SVM 不存在不可分区域，提高了分类精度;

2) 需要构造的分类器少，对于M分类问题，仅需要构造$M-1$个分类器;

3) 从上至下每一层次所需的训练样本及支持向量的数量递减，缩短了训练时间;

4) 分类时不必遍历所有分类器，缩短了分类时间。

其主要缺点是存在误差积累问题，即如果在某个节点上发生分类错误，则错误会沿树结构向后续节点延续，最终导致分类结果与实际情况相去甚远的现象，如图 1 所示，属于类别 1 的样本A在决策点SVM_0处发生错分，这个错误继而被及传递下去，最终导致分类错误。

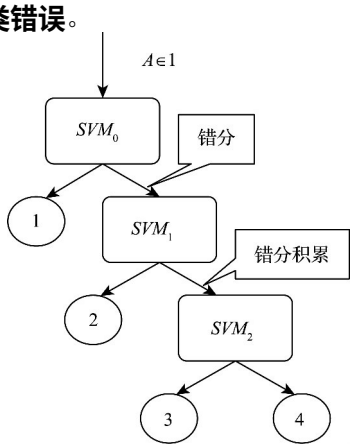

图 1 决策树错分积累

Fig. 1 Decision tree error accumulation

可见,决策树的结构对分类性能影响甚大。如何构造性能良好的决策树结构,尽可能减少误差积累,是本文研究多分类支持向量机所要解决的关键问题。

3　基于粒子群算法的决策树 SVM 多分类方法

如前所述,决策树结构存在误差积累问题,对整体性能影响较大的是上层节点的子分类器,所以在分类过程中应遵循从易到难的原则,首先分割容易分离的类,再到较难分的类,使分类错误尽可能远离根节点,从而得到性能优良的分类器。

3.1　基于粒子群聚类的决策树生成算法

本文利用自变异粒子群(PSO)聚类算法在每个节点将多类训练样本合并为两类,直到叶子节点为止,使两个子类之间的可分性尽可能强,生成最优二叉决策树结构,最终依据该结构来训练子分类器,基于自变异 PSO 聚类的决策树 SVM 生成算法具体流程如下:

Step1: 将全部训练样本集作为初始根节点,在根节点调用自变异 PSO 聚类算法,将原始训练样本合并划分为两类,形成两个子节点。

Step2: 判断子节点是否是包含一个类,若是转向 Step4,若不是则转向 Step3。

Step3: 对该子节点继续调用自变异 PSO 聚类算法,将其再划分为两个子节点,转 Step2。

Step4: 该节点为叶子节点,算法结束。

该方法是在 SVM 训练前针对每个子节点进行的二分类划分,确定各个子分类器对应的位置以及训练样本。

需要说明的是,引入 PSO 聚类算法会延长 SVM 的训练时间,但是在实际应用中,决策树的结构优化和训练过程往往是个离线的过程,以离线过程的耗时增加为代价来换取更优良的 SVM 分类性能是值得且合理的。

3.2　自变异粒子群聚类算法设计

粒子群算法(PSO)[9]是一种有效的基于群体智能理论的全局寻优算法,利用 PSO 算法进行聚类的思想是将聚类视为一种优化问题,在全局范围内利用 PSO 算法得到一个针对数据集的近似最优划分。

PSO 聚类算法需要预先设定簇的个数,一个粒子代表各簇的聚类中心,粒子 X_i 构造如下:

$$X_i = (C_{i1}, C_{i2}, \cdots, C_{ij}) \tag{7}$$

式中: C_{ij} 表示第 i 个粒子所代表的第 j 个类的聚类中心,则每个粒子代表一种对数据集的划分,整个粒子群代表了对数据集的多种划分方案。

PSO 算法的粒子适应度函数为 f。

$$f = \frac{1}{j_e} \tag{8}$$

$$j_e = \sum_{j=1}^{N_c} \sum_{P_m \in C_{ij}} \| P_m - C_{ij} \|^2 \tag{9}$$

式中: j_e 为类内离散度之和, N_c 为簇的个数,数据 P_m 属于聚类中心 C_{ij} 代表的类。可以看出,适应度越高的粒子的类内离散度之和越小,即类内的相似度越高。

在该文中,需要将数据划分为 2 类,则 $N_c = 2$,设 PSO 参数 $C_1 = C_2 = 1.5$,且由于粒子群优化算法存在早熟收敛现象,有可能陷入局部最优解,所以本文算法在迭代中使每个粒子存在 20% 的几率变异为随机粒子,实现全局寻优。自变异 PSO 聚类算法流程描述如下:

Step1: 随机初始化粒子的速度和位置(聚类中心)。

Step2: 按照最邻近法则对数据进行划分,依照适应度的计算公式,计算每个粒子的适应度值,更新个体极值。

Step3: 粒子 20% 可能性发生变异,并寻找全局极值和全局极值位置。

Step4: 按粒子群算法的位置公式和速度公式更新粒子的位置及速度。

Step5: 若达到结束条件,输出最优粒子的位置即最优的 2 个聚类中心; 若未达到结束条件,则返回 Step2。

算法的结束条件可以是达到预设的迭代次数、聚类中心不变(变化很小)或者是簇的成员不再变化。

4　实验结果与分析

在实验中,选取合肥工业大学智能检测实验室的人群密度分类图片库,Polus,A 将人群密度分为 5 个类别(参照表 1),每幅图片只属于其中一个类别。

表 1 人群密度级别

Table1 Crowd density level

密度级别	密度描述	密度范围/(m^2)
1	低密度	0 ~ 0.75
2	较低密度	0.8 ~ 1.5
3	较密集	1.55 ~ 2.5
4	密集	2.55 ~ 3
5	拥堵	3

每幅图像经过预处理后，提取灰度共生矩阵，方向参数 θ 取 0°和 90°，选取能量、熵、对比度、相关性 4 个纹理特征作为图像分类特征，每个人群密度级别分别取 60 张图片，其中 30 张作为训练样本，30 张作为测试样本。实验平台为 intel core2 Duo 2.1 GHz，2GRAM，操作系统为 WIN7。选取径向基函数作为 SVM 核函数，用本文算法进行训练和测试，得到的分类结果如图 2 所示。

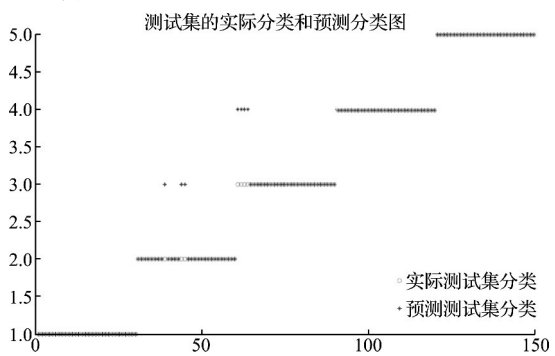

图 2 PSODT-SVM 分类结果

Fig. 2 PSODT-SVM classification result

实验表明，本文的 PSODT-SVM 算法具有较高的分类精度和较快的分类速度，并且提高了 SVM 多分类器的推广效果，取得了优良的分类性能。

使用本文算法与 1-a-1、1-a-r、DT-SVM(随机结构) 3 种方法在 UCI 数据库的 Optdigits 数据集上做对比试验，该数据集包含 10 个类别，特征维数为 64，取 3 820 个样本作为训练集，1 800 个样本作为测试集，实验结果如表 2 所示。

表 2 PSODT-SVM 与 1-a-1、1-a-r、DT-SVM 的性能比较

Table2 Performance Comparison of four algorithms

性能指标	1-a-1	1-a-r	DT-SVM	PSODT-SVM
所需训练 SVM 数量	45	10	9	9
训练时间/s	27.3	31.2	23.6	28.7
分类时间/s	11.3	10.7	8.1	7.9
分类精度(%)	94.3	94.85	93.18	95.53

实验结果表明文本算法在需要训练的 SVM 数量上、分类精度及分类时间上优于 1-a-1、1-a-r、DT-SVM 方法，但需要的训练时间较长，略高于 1-a-1 和 DT-SVM 方法，低于 1-a-r 方法。

5 结 论

本文提出了一种基于粒子群聚类的决策树 SVM 多分类优化算法，用其构建的 SVM 多分类器对不同人群密度级别的图片进行了分类，实验结果表明，该方法提高了分类精度，在一定程度上缩短了分类时间，具有一定实用价值。由于粒子群聚类算法的引入，相对的训练时间会有所延长，但是在实际应用中，决策树的结构优化和训练过程往往是个离线的过程，以离线过程的耗时增加为代价来换取更优良的 SVM 分类性能是值得且合理的。

参考文献

[1] 周绍磊，廖剑，史贤俊. RBF-SVM 的核参数选择方法及其在故障诊断中的应用 [J]. 电子测量与仪器学报，2014，28(3)：240-246.
ZHOU SH L，LIAO J，SHI J X. Kernel parameter selection of RBM-SVM and its application in fault diagnosis [J]. Journal of Electronic Measurement and Instrument，2014，28(3)：240-246.

[2] 魏星. 基于 SVM 的山体滑坡灾害图像识别方法 [J]. 电子测量技术，2013，36(8)：65-70.
WEI X. Landslide disaster image recognition algorithm based on SVM [J]. Electronic Measurement Technology，2013，36(8)：65-70.

[3] 刘松松，张辉，毛征，等. 基于 HRM 特征提取和 SVM 的目标检测方法 [J]. 国外电子测量技术，2014，33(10)：38-41.
LIU S S，ZHANG H，MAO ZH，et al. Target detection method based on HRM extracting and SVM [J]. Foreign Electronic Measurement Technology，2014，33(10)：38-41.

[4] HSU C，LIN C. A comparison of methods for multiclass support vector machines [J]. Neural Networks，IEEE Transactions on，2002，13(2)：415-425.

[5] 唐发明，王仲东，陈绵云. 支持向量机多类分类算法研究 [J]. 控制与决策，2005，20(7)：746-749.
TANG F M，WANG ZH D，CHEN M Y. On multiclass classification methods for support vector machines [J]. Control and Decision，2005，20(7)：746-749.

[6] TAKAHASHI F，ABE S. Decision-tree-based multiclass

support vector machines [C]. Proceedings of the 9th International Conference on Neural Information Processing, ICONIP'02, IEEE, 2002: 1418-1422.

[7] 连可,陈世杰,周建明,等. 基于遗传算法的 SVM 多分类决策树优化算法研究 [J]. 控制与决策,2009(1): 7-12.
LIAN K, CHEN SH J, ZHOU J M, et al. Study on a GA-based SVM decision-tree multi-classification strategy [J]. Control and Decision, 2009(1): 7-12.

[8] MADZAROV G, GJORGJEVIKJ D, CHORBEV I. A Multi-class SVM classifier utilizing binary decision tree [J]. Informatica (Slovenia), 2009, 33(2): 225-233.

[9] 余建平,周新民,陈明. 群体智能典型算法研究综述 [J]. 计算机工程与应用,2010,46(25).
YU J P, ZHOU X M, CHEN M. Research on representative algorithms of swarm intelligence [J]. Computer Engineering and Applications, 2010, 46(25).

[10] WANG R, KWONG S, CHEN D, et al. A vector-valued support vector machine model for multiclass problem [J]. Information Sciences, 2013(235): 174-194.

[11] KECSKÊS I, SZÊKÁCS L, FODOR J C, et al. PSO and GA optimization methods comparison on simulation model of a real hexapod robot [C]. 2013 IEEE 9th International Conference on Computational Cybernetics (ICCC), IEEE, 2013: 125-130.

[12] KENNEDY J. Particle swarm optimization [M]. Encyclopedia of Machine Learning. Springer US, 2010: 760-766.

[13] HUANG C L, DUN J F. A distributed PSO-SVM hybrid system with feature selection and parameter optimization [J]. Applied Soft Computing, 2008, 8(4): 1381-1391.

[14] 陈仁文,朱霞,徐栋霞,等. 基于改进型粒子群算法的卡箍直径检测算法研究 [J]. 仪器仪表学报,2014,35(8):1837-1843.
CHEN R W, ZHU X, XU D X, et al. Spring clamp diameter detection algorithm based on the improved particle swarm optimization [J]. Chinese Journal of Scientific Instrument, 2014, 35(8): 1837-1843.

[15] RANA S, JASOLA S, KUMAR R. A review on particle swarm optimization algorithms and their applications to data clustering [J]. Artificial Intelligence Review, 2011, 35(3): 211-222.

作者简介

王道明,1989 年出生,本科毕业于合肥工业大学,现为合肥工业大学在读研究生。目前主要研究方向为机器学习、图像处理以及模式识别。
E-mail: wang198927@163.com

Wang Daoming was born in 1989, and graduated from Hefei University of Technology. And he is M. Sc. candidate in Hefei University of Technology now. His present research interests include machine learning, image processing, and pattern recognition.

鲁昌华,1962 年出生,2001 年博士毕业于中国科学院研究生院,现任合肥工业大学教授。目前主要研究方向为智能信息处理。
E-mail: lch6208@163.com

Lu Changhua was born in 1962, and graduated from University of Chinese Academy of Sciences. And he is professor in Hefei University of Technology now. His present research interests include intelligent information processing.

蒋薇薇,1978 年出生,现为合肥工业大学在读博士研究生、讲师。目前主要研究方向为智能信息处理。
E-mail: cttjww@126.com

Jiang Weiwei born in 1978, graduated from Hefei University of Technology, Ph. D. candidate in Hefei University of Technology. His present research interests include intelligent information processing.