DOI:10.11721/cqnuj20160327

基于二叉树支持向量机多类分类算法的研究*

吴恩英，吕　佳
(重庆师范大学 计算机与信息科学学院，重庆 401331)

摘要:基于二叉树的支持向量机多类分类算法虽然在目前现有的多类分类算法中总体性能较优,但是仍然存在分类精度和分类效率不高的问题。针对这些问题,提出了一种新的基于欧氏距离的二叉树支持向量机(Distance binary tree SVM,简称 DBT-SVM)多类分类算法,该算法综合地考虑了两类最近样本的欧式距离、类中心的欧氏距离对分类的影响,并且使最容易分离的类能优先分离出来。通过在 UCI 标准数据集上进行实验验证,结果表明该算法行之有效。

关键词:支持向量机;多类分类;二叉树;欧氏距离
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:1672-6693(2016)03-0102-05

支持向量机(Support vector machines,SVMs),是一种建立在统计学习理论的 VC 维理论和结构风险最小化原则基础上的一种有监督的机器学习方法[1]。它比其他学习方法在解决小样本、非线性和高维模式识别中具有更好的特性[2]。支持向量机最初是应用在模式识别中,后被广泛应用到人脸的检验、图像的处理(包括图像的过滤、分类和检索等)和语音识别等方面并取得不错的效果,而这涉及到多类分类问题,因此将支持向量机[3-4]的优势应用到多类分类问题中已经成为众学者关注的焦点。

目前,对于多类分类问题,利用 SVM 解决的方式[5]有两种:一种是“分解-重组”法,即将多类分类的问题分解成多个二类分类的问题,再用 SVM 二值分类器重新组合来解决多类分类问题;另一种是“直接求解”法,即将多类分类问题作为一个整体求解,该方法只需要构造一个分类器便能解决多类分类问题。由于“直接求解”法所构造的分类器结构通常比较复杂,训练时间长,因此第一种方法更受到学者们的青睐。目前构造多个 SVM 二值分类器的方法主要包括一对一 SVM 多类分类法、一对多 SVM 多类分类法、有向无环图 SVM 多类分类法和二叉树 SVM 多类分类法等方法。在以上几种方法中,二叉树多类分类法具有需要训练的二类分类器少的优点,应用更广泛,本文在对其简单介绍的基础上,针对生成的二叉树结构不完全的问题,提出了一种改进的二叉树 SVM 多类分类算法用于解决多类分类问题。

1 二叉树 SVM 多类分类算法

二叉树 SVM 多类分类算法是研究者受“二叉树”思想的启发,将其与 SVM 相结合成二叉树 SVM(Binary tree SVM,简称 BT-SVM)多类分类算法。该算法的思想大概是:先将所有样本的类别分成两个子类,然后将子类进一步分成两个次子类,如此循环下去,直到所有的结点都只包含一个单独的类别为止,然后在每一个非叶子节点处训练一个二值 SVM 的分类器,当且仅当二叉树的结构接近正态时,此时的训练速度和分类的精度才是最佳的[6]。图 1 是常见的两种二叉树结构。

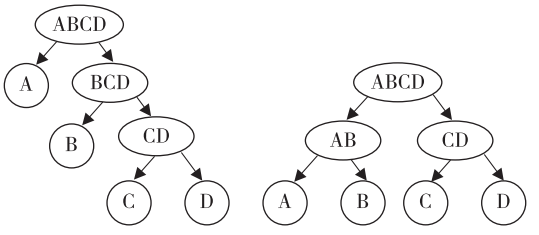


图 1　常见的两种二叉树结构

Fig. 1　Two kinds of two tree structure in common

由图 1 可知,面对相同的分类问题会得到不同的二叉树结构,所以就会产生不同的分类模型,另外对于上层

* **收稿日期**:2016-01-07　**网络出版时间**:2016-04-29　18:34
资助项目:重庆市自然科学基金(No. cstc2014jcyjA40011);重庆市教育委员会科学技术项目(No. KJ1400513)
作者简介:吴恩英,女,研究方向为现代教育技术,E-mail:87534480@qq.com;**通信作者**:吕佳,教授,E-mail:lvjia@cqnu.edu.cn
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.n.20160429.1834.010.html

的结点一旦分类出错，这种错误就会继续传递下去，导致后续结点分类失去意义，造成“错误累积”现象[7]。

针对以上二叉树 SVM 分类算法存在的问题，已经有学者对二叉树 SVM 分类算法提出了改进，目前常见的改进算法思想如下：

1）类间距离法[8]：该方法通过分析已知类别样本的先验分布知识，构造一个二叉决策树，使容易区分的类别从根结点开始逐层被分离出来，从而获得较高的推广能力。

2）双支持向量机法[9]：该方法结合二叉树支持向量机与双支持向量机的优点，利用了二叉树支持向量机解决了传统方法存在的不可分区域问题，并利用双支持向量机克服了一对其余支持向量机多类分类方法的样本不平衡问题，同时利用了双支持向量机的快速优点进一步加快了二叉树支持向量机的训练速度，从而缩短了分类器的训练时间。

3）超球体（或超长方体）最小类包含法[10-11]：该算法所使用的样本分布度量是超球体的半径或者超长方体的体积，目的是比较各个类别分布范围的相对大小，使分布广（即半径大或者体积大）的类别在属性空间中获取更大的划分区域，以此提高多分类模型的推广性。

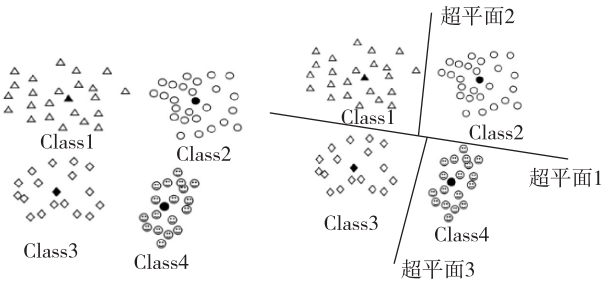
4）动态剪枝法[12]：该算法在分类的时候会去掉一些相对没有价值的支持向量，根据类间的相似度重新构造二叉树，然后剪掉没有价值的枝节，以减少支持向量的个数，缩减训练的时间，加速分类的过程。

以上几种二叉树 SVM 的改进算法的实验性能在一定程度上相对二叉树 SVM 多类分类算法的实验性能都有所提高，但是每种算法在实际分类过程中都很难生成近似完全或者完全二叉树结构，致使分类精度和分类效率仍有待提高，本文针对这些问题，提出了一种新的基于欧氏距离的二叉树 SVM（Distance binary tree SVM，简称 DBT-SVM）多类分类算法。

2 DBT-SVM 多类分类算法

目前，基于二叉树 SVM 多类分类算法所使用的两种二叉树结构（图 1）中，偏二叉树结构的节点都是单类对多类的分割，分类的时候在每个节点都能识别出一个类别；而完全二叉树结构的各节点是对包含的类别等分或者近似的等分，也就是说左右子节点样本类别数目近似相同[13]。虽然偏二叉树结构算法生成简单，但是对于同样类别数目的样本来说树的深度要大，训练时间长，整体的泛化能力弱。由于二叉树的结构直接关系到整个算法的性能，因此如何使二叉树的结构趋向更加理想的状态，让比较容易分离的类别在上层结点处更早地被分离出来，最终使构造出来的超平面获得更加理想的分类性能是本文改进算法所达到的要求。

DBT-SVM 多类分类算法利用了近似完全二叉树的生成策略以及聚类中的类距离的相关定义（定义 1～ 3），在决策结点处以多个类与多个类分割的方式将多类别样本分成两组，使得每组样本数据的分歧最小，并且让两组样本的聚类中心距离最大，从而使生成的二叉树达到近似完全甚至完全的结构，并使得更容易分割的类在上层结点处最先分离。图 2（a）是 4 类样本的二维空间分布图，从直观上看，Class1 和 Class4 相离最远最易分离，Class2 和 Class3 分别聚在 Class1 和 Class4 中，然后 Class1 和 Class2、Class3 和 Class4 分别分离即可，这样既能保证二叉树的结构达到理想的完全状态又能优先分离容易分离的类。



a　各类样本分布　　b　分割图

图 2　DBT-SVM **多类分类算法示意图**

Fig. 2　Schematic diagram of DBT-SVM multi class classification algorithm

设 X 是包含 k 个类别的样本集，X_i 表示第 i 类的训练集，现有如下定义：

定义 1　类 i 与类 j 之间最近样本之间的欧氏距离[14]为：

$$d_{i,j}=\min\{\|x_i-x_j\|\},\tag{1}$$

其中，$i=1,2,\cdots,k$，则有 $d_{i,i}=0$，$d_{i,j}=d_{j,i}$。

定义 2　第 i 类的样本集的中心[8]为：

$$c_i=\frac{1}{n}\sum_{x\in X_i}x,\tag{2}$$

其中,n 是第 i 类的样本总数。

定义 3 若 c_i 和 c_j 分别是类 i 和类 j 的样本中心,则类 i 和类 j 之间的样本中心欧氏距离为:

$$d'_{i,j}=\| c_i-c_j \| 。 \tag{3}$$

DBT-SVM 多类分类算法步骤如下:

步骤 1,首先将类标号按照从小到大的顺序存入集合 S 中,然后根据以上介绍的 3 个定义构造类 i 和类 j 之间的欧式距离矩阵 $\boldsymbol{D}$,其中矩阵的第 1 列代表类 i 的类标号,第 2 列代表类 j 的类标号,第 3 列代表类 i 和类 j 之间最近样本的欧氏距离 $d_{i,j}$,第 4 列为类 i 和类 j 之间样本中心的欧式距离 $d'_{i,j}$。如果是 k 类分类问题,则矩阵 $\boldsymbol{D}$ 如下所示:

$$\boldsymbol{D}=\begin{bmatrix} 1 & 2 & d_{1,2} & d'_{1,2} \\ 1 & 3 & d_{1,3} & d'_{1,3} \\ \vdots & \vdots & \vdots & \vdots \\ k-1 & k & d_{k(k-1)} & d'_{k(k-1)} \end{bmatrix}。 \tag{4}$$

步骤 2,在矩阵 $\boldsymbol{D}$ 中找到集合 S 中拥有最近样本间距离最大($\max d_{i,j}$)的两个类 i 和 j,并按大小存入 $S1$ 和 $S2$ 中,此时在二叉树的结点处训练一个 SVM 分类器,并且更新 $S=S-(S1\cup S2)$;

步骤 3,若 $S=\varphi$,则转到步骤 5;

步骤 4,在 $\boldsymbol{D}$ 中查找类 m(其中 $m\in S$)分别到 $S1$,$S2$ 中各类的最小类中心欧式距离 $d'_{i,m}$ 和 $d'_{j,m}$,如果 $d'_{i,m}\leqslant d'_{j,m}$,则将类别 m 加入集合 $S1$ 中,否则加入集合 $S2$,并转至步骤 2。

步骤 5,分别将集合 $S1$ 和集合 $S2$ 作为二叉树的左右子树,完成一次二类分类。

步骤 6,更新 $S=S1$,直到每个子树只包含一个类别不可再分为止,此时这个类别作为二叉树的叶子结点,算法结束。否则,转至步骤 2,将左子树进一步分割成 2 个次子树。

步骤 7,更新 $S=S2$,直到每个子树只包含一个类别不可再分为止,此时这个类别作为二叉树的叶子结点,算法结束。否则,转至步骤 2,将右子树进一步分割成 2 个次子树。

3 仿真实验

为了测试本文提出的 DBT-SVM 多类分类算法的性能,将其应用在 UCI[15] 数据库中的 back up,iris 和 zoo 等 3 个数据集上(表 1),对算法的分类精度和分类效率进行了仿真实验,并采用一对一 SVM(One-versus-one,简称 OVO)算法、一对多 SVM(One-versus-rest,简称 OVR)算法和 BT-SVM 算法进行对比实验。

表 1 实验数据集信息

Tab. 1 Experimental data set information

数据集	样本数/个	属性个数/个	类别数/个
Back up	305	36	18
iris	150	5	3
zoo	101	17	7

实验中参数的设置如下:核函数采用径向基核函数 $K(x_i,x_j)=\exp\left(\dfrac{-\gamma\| x_i-x_j \|^2}{2\gamma^2}\right)$,惩罚参数 C 和核函数参数 γ 采用网格搜索法来选择,利用 k 折交叉验证找到这对参数的最佳组合,最终确定参数 C 和 γ 的范围是 $(2^{-5},2^{10})$。算法用 Matlab 7.0 实现,实验结果分别见图 3 和图 4 所示。

图 3 是数据集 Back up,iris 和 zoo 分别通过 OVO 算法、OVR 算法、BT-SVM 算法和 DBT-SVM 算法的分类识别率对比直方图,通过对比可以看出,DBT-SVM 算法在 3 个数据集上的分类准确率均高于对比算法,其中在数据集 Back up 上 DBT-SVM 算法比 BT-SVM 算法识别率提高 6.8%。图 4 所示为数据集 Back up、iris 和 zoo 分别通过 DBT-SVM 算法、BT-SVM 算法、OVR 算法和 OVO 算法在分类训练时间的对比直方图,通过比较可以得出,DBT-SVM 算法在 3 个数据集上训练分类时间均比对比算法缩短,尤其是在数据集 Back up 上使用 OVO 算法训练时间为 1.74 s,而使用 DBT-SVM 算法仅用 0.71 s 即可完成样本的分类训练过程。另外由于 OVO 算法在训练阶段需要计算所有的二值分类器,所用的时间更长一些,而基于二叉树的支持向量机算法在训练二值分类器时并不需要利用所有的训练样本。总之,从识别率和分类训练时间两方面比较得出,本文的改进思想有效可行。

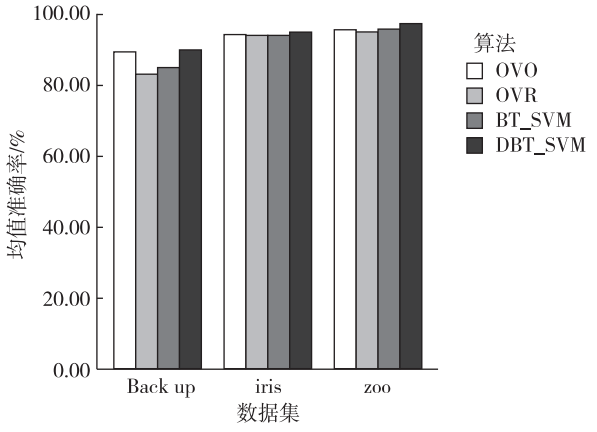


图 3　4 种算法的准确率比较 4 种算法的准确率比较

Fig. 3　Comparison of accuracy of the 4 algorithms

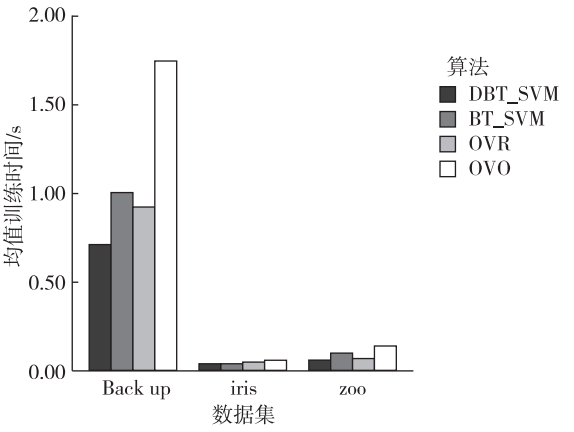


图 4　4 种算法的分类时间

Fig. 4　The classification time of 4 algorithms

4 结语

本文首先研究了现有的二叉树支持向量机算法在解决多类分类问题时，由于生成的二叉树结构基本上是偏二叉树结构，极易导致分类精度和分类效率不高的问题出现，然后针对此问题提出了 DBT-SVM 多类分类算法，利用近似完全二叉树生成策略和聚类中的类距离的相关定义，来解决二叉树结构不完全的问题。改进后的二叉树在构造过程中，遵循最容易分离的类最先分离出来的原则，并且能够保证二叉树的结构达到理想状态。最终的实验结果表明，DBT-SVM 多类分类算法与 OVO 多类分类算法、OVR 多类分类算法和 BT-SVM 多类分类算法相比识别率有所提高，分类训练时间明显缩短。将改进后的二叉树支持向量机应用于实际问题中有很大的研究价值，该算法思想应用于教学质量评价中是下一步的研究方向。

参考文献：

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

[2] 李燕玲. BT-SVM 多分类算法在教学质量评价中的应用[D]. 南宁：广西大学，2014.
Li Y L. Application of BT-SVM multi classification algorithm in the evaluation of teaching quality[D]. Nanning：Guangxi University，2014.

[3] 王华秋，王斌. 优化的邻近支持向量机在图像检索中的应用[J]. 重庆理工大学学报：自然科学版，2014(9)：66-71.
Wang H Q，Wang B. Application of optimized proximal support rector machine in image retrieval[J]. Journal of Chongqing University of Technology：Natural Science，2014(9)：66-71

[4] 张兢，杨超，曾建梅，等. 基于遗传算法与支持向量机的 EMD 改进算法[J]. 重庆理工大学学报：自然科学版，2015(11)：101-105.
Zhang J，Yang C，Zeng J M，et al. Improved EMD Method based on Genetic algorithm and support vector machine [J]. Journal of Chongqing University of Technology：Natural Science，2015(11)：101-105.

[5] 宋召青，陈垚. 基于支持向量机的多类分类算法综述[J]. 海军航空工程学院学报，2015，30(5)：442-446.
Song Z Q，Chen Y. The review of the multi class classification algorithm based on SVM[J]. Journal of Naval Aeronautical Engineering Institute，2015，30(5)：442-446.

[6] 李燕玲，苏一丹. 改进的二叉树支持向量机在多分类中的应用[J]. 计算机技术与发展，2014(7)：181-184.
Li Y L，Su Y D. Improved binary tree support vector machine in multi classification application[J]. Computer technology and development，2014(7)：181-184.

[7] 赵亮. 一种改进的基于支持向量机的多类分类方法[J]. 计算机应用与软件，2014，31(12)：233-236.
Zhao L. An improved multi class classification method based on support vector machine[J]. Computer Applications and Software，2014，31(12)：233-236.

[8] 夏思宇，潘泓，金立左. 非平衡二叉树多类支持向量机分类方法[J]. 计算机工程与应用，2009，45(17)：167-169.
Xia S Y，Pan H，Jin L Z. Non balanced binary tree multi class support vector machine classification method [J]. Computer Eengineering and Applications，2009，45(17)：167-169.

[9] 谢娟英，张兵权，汪万紫. 基于双支持向量机的偏二叉树多类分类算法[J]. 南京大学学报：自然科学版，2011，47(4)：354-363.
Xie J Y，Zhang B Q，Wang W Z. Multi class classification algorithm based on double support vector machine[J].

Journal of Nanjing University: Natural Science Edition, 2011,47(4):354-363.

[10] 黄扬帆,张慧敏,徐子航,等.超球体支持向量机的不完全二叉树多类分类算法[J].重庆大学学报:自然科学版,2012,35(6):125-128+140.
Huang Y F, Zhang H M, Xu Z H, et al. A multi class classification algorithm based on the incomplete two fork tree of the super sphere support vector machine[J]. Journal of Chongqing University: Natural Science Edition, 2012, 35(6):125-128+140.

[11] 刘健,刘忠,熊鹰.改进的二叉树支持向量机多类分类算法研究[J].计算机工程与应用,2010,46(33):117-120.
Liu J, Liu Z, Xiong Y. An improved two cross tree support vector machine multi class classification algorithm research[J]. Computer Engineering and Application, 2010, 46(33):117-120.

[12] 郑伟,马楠.一种改进的决策树后剪枝算法[J].计算机与数字工程,2015(6):960-966+971.
Zheng W, Ma N. An improved decision tree after pruning algorithm[J]. Computer and Digital Engineering, 2015(6):960-966+971.

[13] 范柏超,王建宇,薄煜明.结合特征选择的二叉树SVM多分类算法[J].计算机工程与设计,2010,31(12):2823-2825.
Fan B C, Wang J Y, Bao Y M. Combined with feature selection of binary tree SVM multi classification algorithm [J]. Computer Engineering and Design, 2010, 31(12): 2823-2825.

[14] 唐发明,王仲东,陈绵云.一种新的二叉树多类支持向量机算法[J].计算机工程与应用,2005,41(7):24-26.
Tang F M, Wang Z D, Chen M Y. A new two cross tree support vector machine algorithm[J]. Computer Engineering and Application, 2005, 41(7):24-26.

[15] UCI Repository of Machine Learning Databases and Domain Theories[EB/OL]. (2015-12-20). ftp://ftp.ics.uci.edu/pub/machine-learning-databases.

Research on Multiclass Classification Algorithm Based on Binary Tree SVM

WU Enying, LÜ Jia

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

Abstract: Multiclass classification algorithm based on binary tree SVM has good performance in many multiclass classification algorithms, algorithms; however, its classification accuracy and efficiency are still not high. To solve these problems, a new binary tree SVM multiclass classification algorithm based on euclideanEuclidean distance is proposed in this paper. The algorithm considers the effect of the euclideanEuclidean distances between the two nearest samples and two class centers. Moreover, the algorithm can give priority to those easiest separated classes. The experimental results on UCI benchmark datasets show that the presented algorithm is effective.

Key words: support vector machines; multiclass classification; binary tree; Euclidean distance

(责任编辑 游中胜)