基于格贴近度的SVM决策树层次结构设计方法

赵成龙,张 冉
ZHAO Cheng-long,ZHANG Ran

泰山职业技术学院 信息工程系,山东 泰安 271000
Department of Information Engineering,Taishan Polytechnical Institute,Taian,Shangdong 271000,China

ZHAO Cheng-long,ZHANG Ran.Method of designing hierarchical structure of SVM decision tree based on lattice close-degree.Computer Engineering and Applications,2008,44(27):154-156.

Abstract: Design the hierarchical structure of SVM decision tree based on the lattice close-degree to settle the fuzzy multi-classification.The experimental results show that the SVM decision tree algorithm based on this hierarchical structure has good separation ability.

Key words: lattice close-degree;Support Vector Machine(SVM);hierarchical structure;decision tree

摘 要:利用格贴近度对模糊集的贴近程度进行度量,给出一种基于格贴近度的SVM决策树层次结构设计方法,从而解决对多类模糊样本的分类问题。实验结果表明:基于该层次结构设计方法得到的多类分类器,对多类模糊样本具有良好的分类效果。

关键词:格贴近度;支持向量机;层次结构;决策树
DOI:10.3778/j.issn.1002-8331.2008.27.049 **文章编号:**1002-8331(2008)27-0154-03 **文献标识码:**A **中图分类号:**TP181

1 引言

支持向量机(Support Vector Machine,SVM)是Vapnik等人[1]提出的一种基于统计学习理论的新一代学习机器,它建立在结构风险最小化原则基础之上,在解决有限样本的数据分类时具有很强的学习能力和泛化能力,SVM目前已成为国内外机器学习领域研究的热点。支持向量机最初是针对两类分类问题提出的,由于在实际应用中,多类分类问题是比较普遍的,因此如何将支持向量机的优良性能有效地推广到多类分类问题中,成为支持向量机研究的热点之一。目前已有的SVM多类分类方法有:一对多[2]、一对一[3]、纠错输出编码[4]、有向无环图支持向量机[5]、SVM决策树[6],以及直接方法等,其中对于类别数目较多的分类问题,SVM决策树是提高识别效率的有效方法。

在实际生活中存在一些模糊多类分类问题(如天气预报,自动分拣信件系统等),或类与类的边界不清晰的多类分类问题,或含有孤立点和噪音数据的问题等多类分类问题,这些问题虽然属于多类分类问题,但无法用传统的SVM多类分类方法来解决,为此,人们提出了模糊支持向量机来解决这类多类分类问题,如:为了突出数据中各个样本点的重要程度的差异,同时也为了减小噪音数据对分类结果的影响,台湾学者Liu Chun-Fu等[7]于2002年提出了一种模糊支持向量机多类分类算法,该算法根据训练数据集中各个样本点对分类结果的影响程度不同,对每一个样本点都给予一个确定该样本点属于某一类的隶属度,从而确定该样本点对分类结果的影响程度,大大提高了支持向量机的泛化能力;李昆仑等[8]于2004年在支持向量机多类分类直接算法中引入模糊成员函数,提出了模糊支持向量机多类分类直接算法QP-MC-FSVM,该算法在处理训练样本数据时,根据它们在训练过程中重要程度的不同,给予不同的隶属度,即有区别对待每个样本点,从而在构造分类超平面时,可以忽略那些对分类结果影响很小的数据,降低了噪音对分类结果的影响,该算法减少了噪音数据或孤立点对分类结果的影响,具有良好的鲁棒性;杨杰[9]于2005年对Chun-Fu Liu模糊支持向量机多类分类算法进行了改进,用类中心点来定义样本的模糊隶属度,该改进算法与原算法相比,具有更高的泛化能力。

本文首先介绍SVM决策树和分离测度,以及模糊集和格贴近度,然后介绍模糊支持向量机,最后给出一种基于格贴近度的SVM决策树层次结构设计方法,并用实验说明基于该层次结构设计方法得到的多类分类器,对多类模糊样本具有良好的分类效果。

2 SVM决策树和分离测度[6]

2.1 SVM决策树

SVM决策树的基本思想是:首先将所有类别分成两个子类,再将子类进一步分成两个次级子类,如此循环下去,直到得到一个单独的类别为止,这样就得到一棵倒立的二叉树,然后对每个决策节点的二类分类问题用SVM解决。SVM决策树方法对于k类分类问题,只需构造$k-1$个SVM分类决策函数,具有较高的分类效率,也不存在拒分区域。

当类别数目较多时,SVM决策树是一种比较有效的分类方法。但是,SVM决策树存在错分累积问题,分类错误在越靠

基金项目:山东省教育厅2006年自然科学科研课题(No.J06P56)。
作者简介:赵成龙(1969-),男,副教授,主要从事数据挖掘、计算机应用等方面的教学与研究。
收稿日期:2007-11-09 **修回日期:**2008-03-07

近树根的地方发生，其分类性能越差。为构造分类性能良好的决策树，可以考虑将容易分（不易产生错分）的类先分离出来，然后再分不容易分的类，这样就能够使可能出现的错分尽可能地远离树根，SVM决策树的层次结构设计是一个非常关键的问题，目前主要是通过使用分离测度来划分各级子类的方法来设计SVM决策树的层次结构。

2.2　分离测度

在设计SVM决策树时，要使每个决策节点的类间隔尽可能地大，首先要根据训练样本集估计各类间的分离测度。所谓类间的分离测度，是对类与类之间的可分程度大小的一个度量，它表示类与类的远离程度。分离测度大，说明这两类比较容易分开。

一般情况下，将类 i 与其余各类间的最小分离测度作为类 i 的分离测度 $sm_i=\min\limits_{\substack{j=1,2,\cdots,k\\ j\neq i}}(sm_{ij})$，即类的分离测度，其中，$sm_{ij}$ 表示类 i 与类 j 之间的分离测度。分离测度最大的类是最易分的类，用 s 表示最易分的类，$s=\arg\min\limits_{i=1,2,\cdots,k}(sm_i)$，$i=1,2,\cdots,k$。

3　模糊集和格贴近度

对于模糊集，通常用格贴近度是度量两个模糊集之间接近程度，下面介绍模糊集和格贴近度概念。

3.1　模糊集

模糊数学是一门应用极为广泛的数学学科，尤其是近10多年来，它与信息科学紧密结合，模糊智能系统，模糊神经网络系统纷纷出现[10-11]。下面给出模糊集合的基本概念：

普通集合论要求：论域 U 中每个元 u，对于集合 $A\subset U$，要么 $u\in A$；要么 $u\notin A$，二者心居其一，且仅居其一，决不允许模棱两可。因而，集合 A 由映射 $C_A:U\to\{0,1\}$ 唯一确定，即集合 A 可由特征函数 $C_A(u)=\begin{cases}1,u\in A\\0,u\notin A\end{cases}$ 来刻划。由于这种函数仅取两个值，所以在表达概念方面有其局限性，即只能表达非此即彼的现象，而不能表达存在于现实中的亦此亦彼的现象。1965年美国计算机与控制论专家查德将普通集合论里特征函数的取值范围由{0，1}推广到闭区间[0，1]，于是便得到模糊集的定义如下。

定义 1　设在论域 U 上给定了一个映射：

$$A:U\to[0,1]$$
$$u\mapsto A(u)$$

则称 A 为 U 上的模糊（Fuzzy）集，$A(u)$ 称 A 的隶属函数（或称为 u 对 A 的隶属度）。

对于某模糊集 A，若 $A(u)$ 仅取0和1两个数时，A 就蜕化为普通集合，即普通集是模糊集的特殊形态。若 $A(u)\equiv 0$，则称 A 为空集 ϕ；若 $A(u)\equiv 1$，则称 A 为全集 U。在给定的论域 U 上可以有多个模糊集，记 U 上的模糊集的全体为 $F(U)$。

3.2　格贴近度

格贴近度是度量两个模糊集之间接近程度的一个数量指标，其中有最大最小贴近度，Hamming贴近度，Euclid贴近度，算术平均最小贴近度，格贴近度等。下面给出模糊集 A、B 在有限论域 $U=\{u_1,\cdots,u_n\}$ 上的内积和格贴近度的定义如下：

定义 2　设 $A,B\in F(U)$，$A=\{(u_1,a_1),(u_2,a_2),\cdots,(u_n,a_n)\}$，$B=\{(u_1,b_1),(u_2,b_2),\cdots,(u_n,b_n)\}$ 称 $A\circ B=\bigvee\limits_{u\in U}(a_i\wedge b_i)$ 为 F 集 A、B 的内积。其中，$\wedge$ 表示取 a_i 与 b_i 中较小的值，$\vee$ 表示取上确界。

定义 3　设 $A,B\in F(U)$，称 $N(A,B)=(A\circ B)\wedge(A^c\circ B^c)$ 为 F 集 A、B 的格贴近度。其中，A^c 为模糊集 A 的补集，且 $A^c(u)=1-A(u)$。

4　模糊SVM

下面以两类模糊样本分类问题为例介绍一下模糊支持向量机的基本概念。假设给定一组带有类别标号以及隶属度的训练样本集 $S=\{(x_1,y_1,\rho_1),\cdots,(x_l,y_l,\rho_l)\}$，每一个训练样本 $x_i\in R^n$ 都给出了与其对应的类别标记 $y_i\in\{-1,1\}$ 以及 x_i 属于该类别的隶属度 ρ_i，$0\leqslant\rho_i\leqslant 1$，$i=1,\cdots,l$。令 $z=\varphi(x)$ 表示特征空间中的向量，而 φ 是由输入空间 R^n 到特征空间 Z 的映射，模糊支持向量机的主要目的是构造一个分类超平面，以分割这两类模糊样本，使得分类间隔最大，由此可以得到下面二次规划问题：

$$\begin{aligned}&\min\ \frac{1}{2}(w\cdot w)+C\sum_{i=1}^{l}\rho_i\xi_i\\&\text{s.t. }\ y_i((w\cdot\varphi(x))+b)\geqslant 1-\xi_i\\&\quad\xi_i\geqslant 0,i=1,\cdots,l\end{aligned}\tag{1}$$

引入拉格朗日乘子 α_i，$i=1,\cdots,l$，得到式(1)对偶问题如下：

$$\begin{aligned}&\min\ \frac{1}{2}\sum_{i,j=1}^{l}y_iy_i\alpha_i\alpha_jK(x_i,y_j)-\sum_{i=1}^{l}\alpha_i\\&\text{s.t. }\ \sum_{i=1}^{l}y_i\alpha_i=0\\&\quad 0\leqslant\alpha_i\leqslant C\rho_i\end{aligned}\tag{2}$$

求解式(2)，可以得到决策函数：$f(x)=\text{sgn}(\sum\limits_{i=1}^{l}\alpha_iy_iK(x,x_i)+b)$。

5　基于格贴近度的SVM决策树层次结构设计方法

假设给定一组带有类别标号和隶属度的训练样本集 $S=\{(x_1,y_1,\rho_1),\cdots,(x_l,y_l,\rho_l)\}$，每个训练样本 $x_i\in R^n$ 都给出了与其对应的类别标记 $y_i\in\{1,\cdots,k\}$ 以及属于该类别的隶属度 ρ_i，$0\leqslant\rho_i\leqslant 1$，$i=1,\cdots,l$。下面给出基于格贴近度设计SVM决策树层次结构的思想，以及基于格贴近度设计层次结构的SVM决策树多类分类器训练算法。首先把训练集 S 按类别分成 k 个子集：$S_1,\cdots,S_k$ 且有 $S=S_1\cup\cdots\cup S_k$，其中，S_i 中的每个元素的类别标号都相同。

基于格贴近度设计SVM决策树层次结构的基本思想：首先计算模糊集 S_i 与 S_j 之间的格贴近度 d_{ij}，$i\neq j$，$i,j=1,2,\cdots,k$，然后计算出模糊集 S_i 与其它所有模糊集之间的格贴近度 $d_i=\max\limits_j(d_{ij})$，$i\neq j$，$i,j=1,2,\cdots,k$，并对它们从小到大进行排序，得到非增序列 $\{d_{i1},d_{i1},\cdots,d_{ik}\}$，其对应的模糊集合为序列为 $\{S_{i1},S_{i1},\cdots,S_{ik}\}$，为了便于算法描述，设在各树结点生成的决策函数是将一类和其余类分开，以4类训练样本为例，给出基于格贴近度设计的SVM决策树层次结构的示意图（图1）。

下面，给出基于格贴近度设计SVM决策树的层次结构的SVM决策树模糊多分类训练算法：

步骤 1　初始化，设 $S=\{S_1,\cdots,S_k\}$，$t=0$。

步骤 2　计算模糊集 S_i 与其它每一个 S_j 之间的格贴近度 d_{ij}，$i\neq j$，$i,j=1,2,\cdots,k$。

步骤 3　计算出模糊集 S_i 与其它所有模糊集之间的格贴近度 $d_i=\max\limits_j(d_{ij})$，$i\neq j$，$i,j=1,2,\cdots,k$，并对它们从小到大进行排

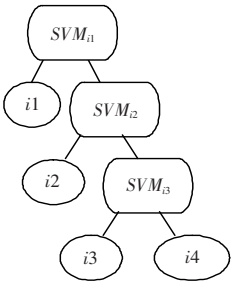

图 1 SVM 决策树示意图

序,得到非增序列$\{d_{i1},d_{i1},\cdots,d_{ik}\}$,其对应的模糊集合为序列为$\{S_{i1},S_{i1},\cdots,S_{ik}\}$。

步骤 4 选择当前 S 中格贴近度最小的模糊类,假设类别标号为 im。

步骤 5 将 S_{im} 与当前 S 中其余所有的模糊类进行分类训练,得到最优判别函数 f_{im},构成树结点。

步骤 6 $S=S-S_{im}$,$t=t+1$。

步骤 7 若 $t<k$,转步骤 3。否则,结束。

6 数值实验

训练样本集 $S=\{(x_1,y_1,\rho_1),\cdots,(x_l,y_l,\rho_l)\}$ 包含 1 000 个样本,即 $l=1\ 000$,$x_i\in R^5$,$y_i\in\{1,2,3\}$,$0\leqslant\rho_i\leqslant 1$。在进行训练前,先对所本样本进行预外理:若 $\rho_i\geqslant 0.5$,则称 x_i 属于类别 y_i 对应的集合。训练时分类器采用径向基核函数,用 3 折交叉验证法估计分类器的正确性,最后得出分类准确率 84.40%。

7 结束语

针对实际生活中存在的模糊分类问题,或类与类的边界不清晰的分类问题,或含有孤立点和噪音数据的分类问题等多类分类问题,研究者们提出了模糊支持向量机。当样本数目比较多时,SVM 决策树是一种有效的识别方法,为了降低 SVM 决策树的错分累积问题,需要有效地设计 SVM 决策树的层次结构。本文利用格贴近度度量模糊集的贴近程度,给出了一种基于格贴近度设计 SVM 决策树的层次结构的方法。实验表明,基于格贴近度的 SVM 决策树对于模糊集具有良好的分类效果。如何针对模糊问题设计出更加合理、有效的多类分类方法,是本文下一步的研究方向。

参考文献:

[1] Vapnik V N.The nature of statistical learning theory[M].New York:Springer,1999.

[2] Weston J,Watkins C.Support vector machines for multi-class pattern recognition[C]//Proceedings of 7th European Symposium on Artificial Neural Networks,Brussels,1999:219-224.

[3] Krebel U.Pairwise classification and support vector machines[C]//Advance in Kernel Methods.Cambridge,MA:MIT Press,1999:255-268.

[4] Dietterich T G,Bakiri G.Solving multi-class learning problem via error-correcting output codes[J].Journal of Artificial Intelligent Research,1995,2:263-286.

[5] Platt J.Large margin DGAs for multi-class classification [C]//Advances in Neural Information Processing System 12.MA:MIT Press,2000:547-553.

[6] Fumitake Takahashi,Shigeo Abe.Decision-tree-based multi-class support vector machines[C]//Proceeding of ICONIP'02.Singapore:IEEE Press,2002:1419-1422.

[7] Lin Chun-fu,Wang Sheng-de.Fuzzy support vector machines[J].IEEE Transactions on Neural Networks,2002,13(2):464-470.

[8] 李昆仑,黄厚宽,田盛丰.模糊多类支持向量机模型[J].电子学报,2004,32(5):830-832.

[9] 杨杰.基于模糊支持向量机的多类分类方法的研究[D].武汉:武汉大学,2005.

[10] Klir G,Yuan B.Fuzzy sets and fuzzy logic:theory and applications[M].Upper Saddle River,NJ:Prentice-Hall,1995.

[11] Zadeh L.Fuzzy sets[J].Inform Control,1965,8:338-353.

[12] 杨伦标,高英仪.模糊数学原理及应用[M].广州:华南理工大学出版,1993.

(上接 133 页)

在网络中的转发次数,进一步提高了能效。仿真结果表明随节点移动速度以及链路差错率的增大,该算法的能耗大大低于传统的盲目泛洪算法,为移动 Ad hoc 网络提供了一种节能的有效途径。

参考文献:

[1] Li L,Halpern J,Bahl P.A cone-based distributed topology-control algorithm for wireless multi-hop networks[J].IEEE/ACM Trans on Networking,2005,13(1):147-159.

[2] Li X Y,Wan P J,Wang Y.Fault tolerant deployment and topology control in wireless networks[C]//The ACM Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc),2003.

[3] Bose P,Gudmundsson J,Smid M.Constructing plane spanners of bounded degree and low weight[C]//The 10th Annual European Symp of Algorithms.London:Springer-Verlag,2002:234-246.

[4] Wattenhofer R,Li L,Bahl P.Distributed topology control for wireless multihop Ad-Hoc networks[C]//The IEEE INFOCOM 2001.Piscataway:IEEE Inc,2001:1370-1379.

[5] Calinescu G.Computing 2-hop neighborhoods in Ad Hoc wireless networks[C]//The Ad-Hoc Networks and Wireless Conf.Berlin:Springer-Verlag,2003:175-186.

[6] Wattenhofer R,Li L,Bahl P.Distributed topology control for power efficient operation in multihop wireless Ad Hoc networks[C]//IEEE INFOCOM 2001.Anchorage:IEEE,2001:1388-1397.

[7] Song W,Wang Y,Li X Y.Localized algorithms for energy efficient topology in wireless Ad hoc networks[C]//The ACM MobiHoc.New York:ACM Press,2004:98-108.

[8] Cartigny J,Simplot D,Stojmenovie I.Localized minimum-energy broadcasting in Ad Hoc networks[C]//The IEEE INFOCOM,2003.San Francisco:IEEE,2003:2210-2217.

[9] Marina M K,Das S R.A topology control approach for utilizing multiple channels in multi-radio wireless mesh networks[C]//The 2nd International Conference on Broadband Networks,2005:381-390.

[10] Huang Z H,Zhang Z S,Ryu B.Impact of topology control on end to end performance for directional manets[C]//Military Communications Conference,2006.Washington D C:IEEE,2006:1-7.

[11] Li N,Hou J C.Topology control in heterogeneous wireless networks:problems and solutions[C]//The IEEE Conf on Computer Communications(INFOCOM).New York:IEEE Press,2004:232-243.

[12] Banerjee S,Misra A.Minimum energy paths for reliable communication in multihop wireless networks[C]//Mobihoc 2002.Lausanne:ACM,2002.