DOI:10.13274/j.cnki.hdzj.2008.04.045

中图分类号:TP301.6 文献标识码:A 文章编号:1009-2552(2008)04-0001-03

基于二叉树的SVM多类分类算法

吕晓丽,李 雷,曹未丰

(南京邮电大学通信与信息工程学院,南京 210003)

摘 要:在介绍了几种常用的支持向量机的多类分类方法及分析其存在的问题和缺点的基础上利用类均值距离思想提出了一种新的基于二叉树的多类SVM分类方法。

关键词:支持向量机;多类分类;类距离;二叉树

SVM multi-class classification based on binary tree

LV Xiao-li, LI Lei, CAO Wei-feng

(School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The problems and defections of the existing methods of SVM multi-class classification were analyzed. A multi-class classification method based on binary tree using class distance was put forward.

Key words: support vector machine (SVM); multi-class classification; class distance; binary tree

0 引言

支持向量机(SVM, Support Vector Machine)是Vapnik[1]在20世纪70代提出并在90年代逐渐完善的一种针对小样本的机器学习论理论。现在支持向量机作为一种分类算法已经得到广泛的应用。但SVM本质上是一种二分类方法在现实世界中大多数分类问题都是多分类问题,因此寻求SVM的多类分类方法,就成了SVM理论与应用研究的一个重要问题。现在应用较广的算法主要有One-versus-one, One-versus-the-rest及有向无环图等算法,但这些算法都存在一定的缺陷,前两种算法会存在大量的不可分区域,且训练时间和测试时间比较长,后一种算法虽然解决了不可分区域问题,但各个分类器在有向无环图中的位置会对整个分类模型的性能产生很大影响。本文在分析了这几种常用的支持向量机多类分类算法存在的问题的基础上利用一种新的类距离思想提出了一种新的基于二叉树的多类SVM分类方法,并与常用的方法进行了比较。

1 支持向量机简介

支持向量机的主要思想是针对两类分类问题寻找一个满足分类要求的最优超平面,使得这个分类超平面在保证分类精度的同时能够使得超平面两侧的空白区域最大化[2];假设训练样本(x_1, y_1), …, (x_l, y_l), $x \in R^n$, $y \in \{+1, -1\}$, l为样本数,分类问题可以描述为:

$$y_i(w \cdot x_i + b) - 1 \geqslant 0 \tag{1}$$

其中,x_i是输入的第i个样本;$y_i \in \{+1, -1\}$表示相应x_i期望分类;w是可调权值向量;b是偏置。

分类间隔为:

$$\gamma = \min_{\{x_i \mid y_i=1\}} \frac{w \cdot x_i + b}{\|w\|_2} - \max_{\{x_i \mid y_i=-1\}} \frac{w \cdot x_i + b}{\|w\|_2} = \frac{2}{\|w\|_2} \tag{2}$$

分类问题的求解就是在条件式(1)下最大化式(2),这是一个典型的二次规划问题,因此目标函数为:

$$\min_w L(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{l} \alpha_i [y_i(w \cdot x_i + b) - 1] \tag{3}$$

其中,$\alpha_i \geqslant 0$是lagrange乘子。

此优化问题可转化为求解其"对偶"问题,即

$$\max_\alpha W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i, x_j) \tag{4}$$

约束条件为:

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{5}$$

收稿日期:2007-09-07

基金项目:国家自然科学基金项目(10371106, 10471114);江苏省高校自然科学基金项目(04KJB110097)

作者简介:吕晓丽(1982-),女,南京邮电大学通信与信息工程学院硕士研究生,主要研究方向为智能信号的处理及应用。

求解二次规划问题式(4)、式(5),Karush-Kuhn-Tucher 条件将起重要的作用,其解必满足:

$$\alpha_i\{[(w\cdot x_i)+b]y_i-1\}=0,\ i=1,2,\cdots,l \quad (6)$$

可见,那些 $\alpha_i=0$ 的样本对于分类问题不起什么作用,只有 $\alpha_i\neq 0$ 的样本才会对最优解 w^*, b^* 起作用,从而决定分类结果。这样的样本定义为支持向量。

对于非线性问题,由泛函有关理论,只要一种核函数 $K(x_i, x_j)$ 满足 Mercer 条件,它就对应某一空间中的内积[3]。因此,在最优分类面中采用适当的核函数就可以实现某一非线性变换后的线性分类,而计算的复杂度却没有增加。此时的目标函数式(4)变为:

$$\max_{\alpha} W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \quad (7)$$

训练后的相应分类函数为:

$$d(x)=\sum_{i=1}^{l}y_i\alpha_i K(x, x_i)+b^* \quad (8)$$

即支持向量机根据 $d(x)$ 的符号来确定输入样本 x 的归属。

2 常用的多类支持向量机分类方法

支持向量机是针对二分类问题的,对于 k 类多分类问题,现在理论研究和工程应用中主要有以下几种多分类方法。

2.1 一对一方法

一对一方法(One-versus-one Method)[4] 是由 Knerr 提出的,该算法在 k 类训练样本中构造所有可能的两类分类器,每类仅仅在 k 类中的 2 类训练样本上训练,结果共构造 $k(k-1)/2$ 个 SVM 子分类器。在构造类 i 和类 j 的 SVM 子分类器时,在样本数据集选取属于类 i 和类 j 的样本数据作为训练样本数据,并将属于类 i 的数据标记为正,将属于类 j 的数据标记为负。测试时,将测试数据对 $k(k-1)/2$ 个 SVM 子分类器分别进行测试,并累计各类的得分,选择得分最高者所对应的类为测试数据的类别。

2.2 一对多方法

一对多方法(One-versus-the-rest Method)[4] 构造 k 个支持向量机子分类器。在构造第 i 个支持向量机子分类器时,将属于第 i 类别的样本数据标记为正类,不属于 i 类别的样本数据标记为负类。测试时,对测试数据分别计算各个分类器的决策函数值,并选取最大函数值对应的类别为测试数据的类别。

2.3 有向无环图 SVM 方法

有向无环图(Directed Acyclic Gsaph)SVM 分类在训练阶段也是采用一对一 SVM 的任意两两组合训练方式,同样也需要构造 C_k^2 个子分类器,但是在分类过程中,DAG 将所用子分类器构造成有向无环图如图 1 所示,包括 C_k^2 个节点和 k 个叶子,其中每个节点是一个子分类器。当对未知样本训练时,从根节点开始分类,只需 $k-1$ 步即可完成分类[5]。

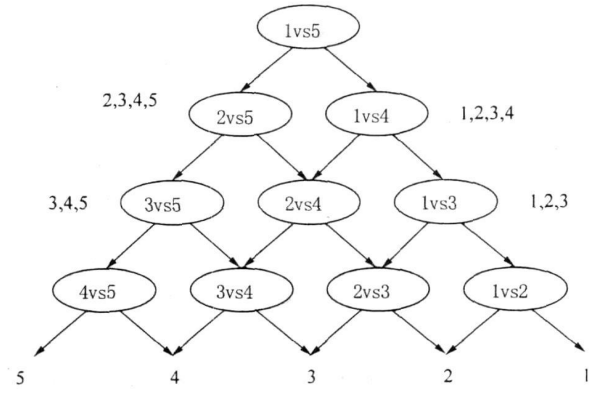

图 1 采用 DAG5 分类决策过程

以上三种方法是最常见的三种分类方法,但存在一定的缺点。前两种存在一个明显的缺陷,当分类系统中的类别较多时,将进行大量的二次规划计算,尤其当样本数量巨大时,还会由于支持向量数目的大量增加而使算法的复杂度急剧增高,系统性能明显下降,DAG SVM 方法和一对一方法相比,在分类过程中减少了重复的操作,大大提高了分类速度。这种分类方法的缺点是未考虑样本不平衡数据对分类速度的影响,而且没有考虑分类错误传递对后续产生的影响。

针对以上几种方法的缺点,本文介绍一种基于二叉树的支持向量机分类方法,并分析其性质。

3 基于二叉树的多类支持向量机分类算法

基于二叉树的多类 SVM 算法是将所有类别分成两个子类,再将两个子类分别划分成两个次级子类,依次类推,直到所有的节点只含有一个类别为止,这样,多分类问题就转化为二分类问题,每个节点处采用 SVM 二值分类器作为分类函数。

二叉树的分类结构主要有两种[6]:一种是在每个内节点处,由一个类与剩下的类构造分割面;另一种是在内节点处,可以是多个类与多个类的分割。本文只考虑前一种情况,即每次分割只分割出一个类。但二叉树的不同的层次结构对分类精度有一定影响,并且这种影响有可能产生“误差累积”现象[7],即若在某个结点上发生分类错误,将会把错误延续下去,该结点后续下一级结点上的分类就失去意义。因此,要合理选择二叉树的层次结构。现有的二叉树生成算法主要有先聚类后分类[8]、聚类分析中的类距离等。类与类之间的距离定义有很多种,例如最短距离法、最长距离法、中间距离法、重心法、类平均距离法等。本文采用一种新的类距离定

义方法作为二叉树的生成算法，基本思想是让与其他类相隔最远的类最先分隔出来，这时构造的最优超平面具有较好的推广性。

首先定义类距离 $\delta_{i,j}$ 为 i 类和 j 类均值向量间的距离减去各自的类的平均半径[9]：

$$\delta_{i,j} = \| m_i - m_j \|^2 - r_i - r_j \tag{9}$$

$$r_i = \frac{1}{n_i}\sum_{k=1}^{n_i} \| x_k^i - m_i \|^2 \tag{10}$$

其中，$m_i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_k^i$ 表示第 i 类样本集的均值向量，$\| m_i - m_j \|^2$ 为 i 类和 j 类均值向量间的距离，r_i 和 r_j 分别为 i 类及 j 类的类平均半径，n_i 为 i 类中的样本数目，称 $\delta_{i,j}$ 为 i 类和 j 类之间的类均值距离，如图2所示。

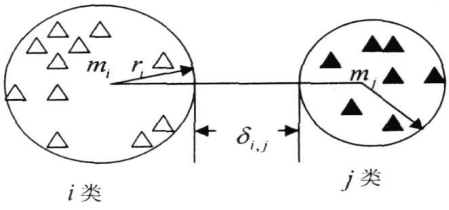

图2　类均值距离 $\delta_{i,j}$

所以，以类均值距离作为二叉树生成算法的多类SVM算法步骤为：

Step1：根据式(9)计算类与类之间的距离 $\delta_{i,j}(i, j = 1, 2, \cdots, k, i \neq j)$。

Step2：对每一类都存在 $k-1$ 个与其他类的距离值，分别对每个类的这些距离值由小到大排列，并重新编号，例如，第 i 类与其他类距离值 $\delta_{i,j}(j = 1, 2, \cdots, k, j \neq i)$ 按由小到大的顺序排列为：$d_i^1 \leqslant d_i^2 \leqslant \cdots \leqslant d_i^{k-1}$。

Step3：根据 $d_i^1(i = 1, 2, \cdots, k)$ 的值由大到小的顺序对相应的类进行排序。当存在两个或两个以上的类具有相同的 d_i^1 时，可比较他们的 d_i^2 大小，如此下去，若 $d_i^1, d_i^2, \cdots, d_i^{k-1}$ 都相同，则把类标号小的排在前面，类标号大的排在后面。最后将得到所有类别的排列 $n_1, n_2, \cdots, n_k$，此时，$n_m \in \{1, 2, \cdots, k\}$，$m = 1, 2, \cdots, k$ 为类标号。

Step4：根据类标号排序，生成如图3所示的二叉树。

Step5：根据生成的二叉树，利用二值SVM训练算法构造二叉树各内节点的最优超平面。以第 n_1 类样本为正样本集，其他样本为负样本集，利用SVM训练算法构造根节点处的二值SVM子分类器。然后删掉第 n_i 类的样本，以第 n_2 类样本为正样本集，第 $n_3, \cdots, n_k$ 类样本为负样本集，构造第二个内节点的二值SVM子分类器。依次下去，直到所有的二值子分类器训练完，从而可得到基于二叉树的多类别

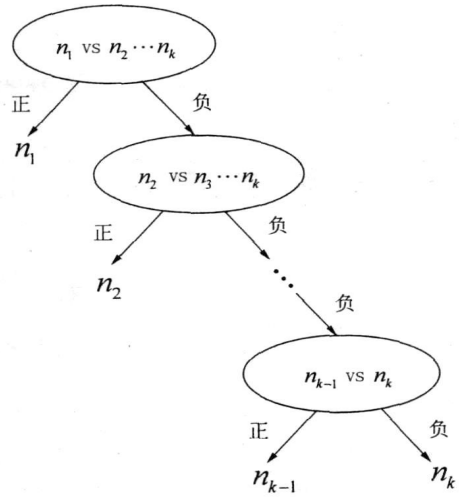

图3　多类SVM分类模型的二叉树结构

SVM分类模型。

Step6：算法结束。

二叉树算法解决了存在不可分盲点、支持向量机数量过多等问题，缩短了训练时间，提高了算法的性能，具有较好的推广能力。

4　结束语

本文在介绍了几种常用的多类SVM分类算法基础上提出了一种新的基于二叉树的多类SVM分类算法，此算法克服了传统算法不可分区域的问题，并且此算法只需构造 $k-1$ 个SVM分类器，测试时也不一定需要计算所有的分类器判别函数，因此可大大节省训练时间。另外此算法利用类距离作为二叉树的生成算法，把与其他类相隔最远的类最先分隔出来，使此算法具有较好的推广能力。

参考文献：

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

[2] Vasehgi S V. State Duration Modeling in Hidden Markov Models[J]. Signal processing, 1995, 41(1): 31－41.

[3] Vladimir N, Vapnik. Statistical Learning Theory[M]. New York: John Wiley&Sons, Inc., 1998.

[4] HSU C2W, L IN C2J. A comparison of methods for multi2class support vector machines[J]. IEEE Transaction on Neural Network, 2002, 13(2): 415－425.

[5] Jaehe Yoon, Azer Bestavros, Ibrahim Adaptive Reliable Multicast[J]. IEEE, 2000, 3(6): 1542－1546.

[6] 唐发明，王仲东，等. 支持向量机多类分类算法研究[J]. 控制与决策，2005, 20(7): 747.

[7] 刘志刚，李德仁，秦前清，等. 支持向量机在多类分类问题中的推广[J]. 计算机工程与应用，2004, 40(7): 10－13.

[8] 孟媛媛，刘希玉. 一种新的基于二叉树的SVM多类分类方法[J]. 计算机应用，2005, 25(11): 2654－2655.

[9] 李海峰，张建州，等. 基于类距离的可分离性判据[J]. 计算机工程与应用，2003, 26: 97－98.

责任编辑：么丽萍