

一种快速的支持向量机多类分类算法

秦玉平¹ 罗倩¹ 王秀坤² 王春立³

(渤海大学信息科学与工程学院 锦州 121000)¹ (大连理工大学电子与信息工程学院 大连 116024)²

(大连海事大学信息科学技术学院 大连 116026)³

摘要 提出了一种快速的支持向量机多类分类算法。首先用每类训练样本的样本数作为权值构造最优二叉树,然后对每个非叶子结点训练两类分类器。分类时,从二叉树根结点开始逐层向下分类,直到某一叶子结点,该结点对应的类别即为待分类样本的类别。在 Reuters 21578 标准数据集上进行的分类实验表明,该算法具有较好的性能,在一定程度上克服了现有的支持向量机多类分类算法分类速度较慢的缺点,尤其在类别数较多、各类样本规模相同的情况下,采用该算法能够较大幅度地提高分类速度。

关键词 支持向量机,多类分类,二叉树

中图法分类号 TP181 文献标识码 A

Fast Multi-class Classification Algorithm of Support Vector Machines

QIN Yu-ping¹ LUO Qian¹ WANG Xiu-kun² WANG Chun-li³

(College of Information Science and Technology, Bohai University, Jinzhou 121000, China)¹

(School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China)²

(College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China)³

Abstract A fast support vector machines multi-class classification was proposed. Firstly, the number of every class training samples is used as weight to construct Huffman binary tree, and then train sub-classifiers for every non leaf node in the binary tree. For the sample to be classified, the sub-classifiers that between the root and a certain left are used to classify, the left is the class of the sample. The classification experiments on the Reuters 21578 database are done with this algorithm. The experimental results show that it has better performance and partly overcomes the flaw of existing multi-class classification algorithm of support vector machines, which is slow in the process of classification. This algorithm can remarkably increase the speed of classification, especially in the case of more classes and the scale of every class is uniform.

Keywords Support vector machine, Multi-class classification, Binary tree

1 引言

支持向量机(Support Vector Machine, SVM)^[1]是在统计学习理论上发展起来的一种新型机器学习方法,在解决小样本、非线性及高维模式识别问题上表现出许多其它机器学习方法不可比拟的优势。支持向量机最初是针对二类分类问题提出的,但在实际生活中更多的还是多类问题。在处理多分类问题时,人们往往将其分解成一系列的二分类问题加以解决。常见的处理方法包括 1-a-1 方法^[2]、1-a-r 方法^[3]、DAGSVM 方法^[4]、纠错编码支持向量机^[5]、二叉树支持向量机^[6,7]、M-ary 支持向量机^[8]和基于线性规划的多类支持向量机^[9]等。其中二叉树支持向量机是一种非常有效的方法,在实际中已得到广泛的应用^[10-12]。但现有的二叉树支持向量

机多类分类方法并非随机地生成二叉树,而是采用二叉树生成算法,因此不能很好地提高整个分类模型的推广能力。另外,它使用了过多的分类器,在一定程度上影响了分类速度。

本文提出了一种新的最优二叉树支持向量机多类分类方法(Huffman Binary Tree SVM, HBT SVM)。用每类训练样本的数量作为权重构造最优二叉树。对于每个非叶子结点,训练支持向量机。对于待分类样本,从二叉树根结点开始逐层向下分类,直到某一叶子结点。本算法在保证分类精度的前提下,有效地提高了分类速度。

本文第2节介绍了支持向量机理论;第3节详细阐述了基于最优二叉树的支持向量机多类分类算法;第4节给出了在 Reuters 21578 标准语料库上的实验结果;最后得出结论。

到稿日期:2009-08-25 返修日期:2009-11-06 本文受国家自然科学基金项目(60603023),国家基础研究重大项目(973)研究专项(2001CCA00700)资助。

秦玉平(1965—),男,博士,教授,主要研究领域为机器学习和决策支持系统, E-mail: jzqinyuping@gmail.com; 罗倩(1985—),女,硕士生,主要研究领域为机器学习; 王秀坤(1945—),女,教授,博士生导师,主要研究领域为数据库系统和决策支持系统; 王春立(1972—),女,博士,教授,主要研究领域为模式识别。

2 支持向量机

设给定训练样本集 $\{x_i, y_i\}_{i=1}^l$ 和核函数 $K(x_i, x_j)$, 其中 $x_i \in R^n, y_i \in \{-1, 1\}$, K 对应某特征空间 Z 中的内积, 即 $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, 变换 $g: X \mapsto Z$ 将样本从输入空间映射到特征空间, 支持向量机的数学模型为:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \quad (1)$$

$$\text{s. t. } y_i(wx_i + b) \geq 1 - \xi_i \quad i=1, \dots, l \quad (2)$$

式中, w 为超平面的法向量, b 为超平面的偏置, ξ_i 是松弛变量, C 为指定的惩罚因子。其对偶问题为:

$$\max \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j K(x_i \cdot x_j) \quad (3)$$

$$\text{s. t. } 0 \leq \alpha_i \leq C \quad (4)$$

$$\sum_{i=1}^l \alpha_i y_i = 0, i=1, \dots, l \quad (5)$$

式中, α 为 Lagrange 乘子。对偶问题的最优解 $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_l\}$ 使得每一个样本 x_i 都要满足优化问题的 KKT 条件:

$$\alpha_i = 0 \Rightarrow y_i f(x_i) \geq 1 \quad (6)$$

$$0 < \alpha_i < C \Rightarrow y_i f(x_i) = 1 \quad (7)$$

$$\alpha_i = C \Rightarrow y_i f(x_i) \leq 1 \quad (8)$$

当 $\alpha_i = 0$ 时, 对应的样本在分类器的间隔以外; 当 $0 < \alpha_i < C$ 时, 对应的样本位于分类间隔之上; 当 $\alpha_i = C$ 时, 位于分类器间隔之内。如果 $0 < \alpha_i \leq C$, 称 x_i 为支持向量, 决策函数 (9) 称为支持向量机。

$$f(x) = \text{sgn}(\sum_{i=1}^l y_i \alpha_i K(x_i, x) + b) \quad (9)$$

3 算法描述

二叉树法(Binary Tree, BT)的思想是先将所有类别划分为两个子类, 每个子类又划分为两个子子类。依此类推, 直到所有结点都只包含一个类别为止。每次划分后, 两类分类问题的规模逐级下降。该方法将原有的多类问题同样分解成了一系列的两类分类问题, 其中两个子类间的分类函数采用 SVM。如何生成二叉树, 是该方法的关键, 因为不同的生成方法直接影响生成的二叉树结构, 从而对整个分类模型的推广能力有较大的影响。现有的二叉树支持向量机多类分类方法大都根据样本分布或类中心距离构造偏置树^[13,14], 分类速度较慢。为此, 提出了一种新的二叉树生成算法。以每类样本的数量作为权重, 按照 Huffman 树的构造过程自下向上地构造二叉树。算法描述如下:

设给定的多类样本集 $A = \{x_i, y_i\}_{i=1}^l$ 和核函数 $K(x_i, x_j)$, 其中 $x_i \in R^n, y_i \in \{1, 2, \dots, N\}$, l 为样本个数, N 为类别数, K 对应某特征空间 Z 中的内积, 即 $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, 变换 $\phi: X \mapsto Z$ 将样本从输入空间映射到特征空间。设第 m 类样本子集 $A^m = \{x_i^m, y_i^m\}_{i=1}^{l_m}$, l_m 为第 m 类样本数。多类分类二叉树生成算法如下。

步骤 1 根据给定的 N 类样本的数量(权值) $\{l_1, l_2, \dots, l_N\}$ 构成 N 棵二叉树的集合 $F = \{T_1, T_2, \dots, T_N\}$, 其中每棵二叉树 T_i 中只有一个权值为 l_i 的根结点, 其左右子树均为空;

步骤 2 在 F 中选取两棵根结点的权值最小的二叉树作为左右子树, 构造一棵新的二叉树。将左子树根结点对应的

样本作为正类, 右子树根结点对应的样本作为负类, 训练支持向量机, 同时置新的二叉树的根结点权值为其左右孩子结点权值之和。

步骤 3 在 F 中删除这两棵子树, 同时将新得到的二叉树加入到 F 中;

步骤 4 重复步骤 2 和步骤 3, 直到 F 中仅剩一棵树为止, 这棵二叉树即为所求。

对待分类样本, 从根结点开始, 依次序经过若干个两类分类器到达某个叶子结点, 该结点对应的类即为该样本的类别。

4 实验结果与分析

实验数据选自标准数据集 Reuters 21578。从中选取 5 类共 896 篇文本, 其中的 598 篇作为训练样本, 其余的 298 篇作为测试样本, 如表 1 所列。

表 1 训练语料和测试语料

类别	wheat	corn	coffee	soybean	cocoa
类别代码	1	2	3	4	5
训练集	204	168	97	79	50
测试集	101	84	48	40	25

将文本数据经过预处理后形成高维词空间向量, 采用信息增益的方法进行特征降维。向量中每个词的权重根据 $TF * IDF$ 公式计算。

实验环境为 CPU Pentium 1.6G, 512M 内存, Windows XP 操作系统。算法采用 RBF 核函数: $K(x, y) = e^{-\gamma \|x-y\|^2}$, 其中, $\gamma = 0.1$ 。系统参数 $C = 10$ 。算法实现参考了 Chang 和 Lin 所开发的 Libsvm^[15], 并在此基础上进行了相应的修改。

实验中采用通用的准确率、召回率和 F_1 值作为评价指标。

$$\text{准确率} = \frac{\text{分类正确的样本数}}{\text{实际分类的样本数}} \quad (10)$$

$$\text{召回率} = \frac{\text{分类正确的样本数}}{\text{应有的样本数}} \quad (11)$$

$$F_1 = \frac{\text{准确率} \times \text{召回率} \times 2}{\text{准确率} + \text{召回率}} \quad (12)$$

式中, 每一类的准确率、召回率和 F_1 值称为微平均; 所有类的准确率、召回率和 F_1 值称为宏平均。

根据本文给出的算法构造如图 1 所示的二叉树。

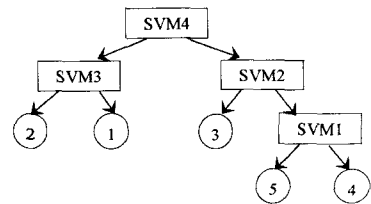


图 1 多类分类二叉树结构

图中, SVM1 是以第 5 类作为正类、以第 4 类作为负类训练得到的分类器; SVM2 是以第 3 类作为正类、以第 4 类和第 5 类作为负类训练得到的分类器; SVM3 是以第 2 类作为正类、以第 1 类作为负类训练得到的分类器; SVM4 是以第 1 类和第 2 类作为正类、以第 3 类、第 4 类和第 5 类作为负类训练得到的分类器。

表 2 给出了算法的宏平均准确率、宏平均召回率和宏平

均 F_1 值的比较。表 3 给出了算法的训练时间和分类时间的比较。

表 2 宏平均准确率、宏平均召回率和宏平均 F_1 值比较

算法	宏平均准确率(%)	宏平均召回率(%)	宏平均 F_1 值(%)
1-a-l	68.27	62.91	64.71
1-a-r	68.51	64.11	65.61
DAGSVM	68.22	62.75	64.23
HBTSVM	69.07	64.15	65.83

表 3 训练时间和分类时间比较

算法	训练时间(ms)	分类时间(ms)
1-a-l	1038	469
1-a-r	2109	422
DAGSVM	1038	357
HBTSVM	986	315

由实验结果可知,在分类精度上,HBTSVM 与 1-a-r 相当,略高于 1-a-l 和 DAGSVM;在训练速度上,HBTSVM 高于 1-a-l, 1-a-r 和 DAGSVM,其原因是 1-a-l 和 DAGSVM 需要训练的子分类器数目较多,虽然 HBTSVM 和 1-a-r 需要训练的子分类器的数目相同,但 1-a-r 的每个子分类器的训练都要求全部训练样本,而 HBTSVM 的子分类器中只有一个要求全部训练样本,其余只要求部分训练样本;在分类速度上,HBTSVM 明显高于 1-a-l,1-a-r 和 DAGSVM,其原因是在对新样本进行分类时,其它 3 种方法要么需要较多的子分类器,要么需要进行较复杂的计算。由此可知,HBTSVM 在保证分类精度的前提下,分类速度有了较大的提高,在应用方面具有一定的优越性。若类别较多、各类训练样本的规模相同,则生成的二叉树为完全二叉树,分类速度的优势更明显。

结束语 多类分类在许多应用领域具有重要的价值。本文对支持向量机多类分类方法进行深入研究,提出了一种新的基于二叉树的支持向量机多类分类算法。在保证分类精度的同时,分类速度有较大提高。采用该算法在 Rutgers 21578 标准数据集上进行了分类实验,结果证明了该算法的有效性。进一步的研究工作是在类别数较多,规模较大的数据集上进行试验。

参 考 文 献

[1] Vapnik V. The Nature of Statistical Learning Theory[M]. New

(上接第 199 页)

结束语 TSP 问题一直是作为检验人工智能算法优劣的标志性问题。本文通过对人工代谢算法的思想来源、编码规则、运行模型、收敛性能地阐述和证明,说明了算法的运作过程,并对 TSP 问题的求解过程进行了详细分析,给出了相关的结论。

参 考 文 献

[1] 王凌. 智能优化算法及其应用[M]. 北京:清华大学出版社,2001
 [2] 陈国良,王煦法,庄镇泉,等. 遗传算法及其应用[M]. 北京:人民邮电出版社,1996
 [3] 焦李成,杜海峰,刘芳,等. 免疫优化计算、学习与识别[M]. 北京:科学出版社,2006
 [4] 丁永生,任立红. 人工免疫系统:理论与应用[J]. 模式识别与人

York, Springer, 1995

[2] Bennett K P. Combining Support Vector and Mathematical Programming Methods for Classification[A]// Advances in Kernel Methods; Support Vector Learning[C]. Cambridge, MA: MIT press, 1999: 307-326
 [3] Krebel U G. Pairwise Classification and Support Vector Machines [A]// Advances in Kernel Methods; Support Vector Learning[C]. Cambridge, MA: MIT press, 1999: 255-268
 [4] Platt J C, Cristianini N, Shawe-Taylor J. Large Margin DAGs for multiclass classification[A]// Advances in Neural Information Processing Systems[C]. Cambridge, MA: MIT Press, 2000: 547-553
 [5] Ghani R. Using error-correcting codes for text classification[C]// Proc of the 17th International Conference on Machine Learning, 2002
 [6] Cheong S. Support vector machines with Binary tree architecture for multi-class classification[J]. Neural Information Processing Letters and Reviews, 2004
 [7] 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): 2653-2657
 [8] 马笑潇,黄席樾,柴毅. 基于 SVM 的二叉树多类分类算法及其在故障诊断中的应用[J]. 控制与决策, 2003, 18(3): 272-276
 [9] 孙德山,吴今培. 基于线性规划的多支持向量机算法[J]. 计算机科学, 2005, 32(10): 160-163
 [10] 朱远平,戴汝为. 基于 SVM 决策树的文本分类器[J]. 模式识别与人工智能, 2005, 18(4): 412-416
 [11] 应伟,王正欧,安全龙. 一种基于改进的支持向量机的多类文本分类方法[J]. 计算机工程, 2006, 32(16): 74-76
 [12] 张永,迟忠先,闫德勤. 一种新的模糊补偿多类支持向量机[J]. 计算机科学, 2006, 33(12): 152-155
 [13] 唐发明,王仲东,陈锦云. 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
 [14] 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
 [15] Chang Chinchang, Lin Chihjen. LIBSVM: a Library for Support Vector Machines[EB/OL]. <http://www.csie.ntu.tw/~cjlin/libsvm>, 2005

工智能, 2000, 13(1): 52-59

[5] 李士勇. 蚁群算法及其应用[M]. 哈尔滨:哈尔滨工业大学出版社, 2004
 [6] 埃伯哈德. O. 沃伊特. 生物化学系统的计算分析[M]. 储炬,李友荣,译. 北京:化学工业出版社, 2006
 [7] 内斯托尔 V. 托雷斯,埃伯哈德. O. 沃伊特. 代谢工程的途径分析与优化[M]. 修志龙,滕虎,译. 北京:化学工业出版社, 2005
 [8] 王翼飞,史定华. 生物信息学——智能化算法及其应用[M]. 北京:化学工业出版社, 2006
 [9] 梅乐和,岑沛霖. 现代酶工程[M]. 北京:化学工业出版社, 2006
 [10] Chekhova E, Barton P I, Gorak A. Optimal operation processes of discrete-continuous biochemical Processes [J]. Computers and Chemical Engineering, 2000, 24: 1167-1173