

基于随机中心距离排序的支持向量预选取方法

胡志军, 王鸿斌, 李 荣

(忻州师范学院 计算机系, 山西 忻州 034000)

摘 要: 提出了一种基于随机中心距离排序的支持向量预选取方法. 对于线性可分情况, 该方法首先从每一个类别中随机选取一定数目的样本计算均值, 并把该均值作为该类别样本的随机中心, 然后对每一个样本计算它与另一类样本随机中心之间的距离, 最后选择一定数目具有较小随机中心距离的原始样本组成边界样本集. 对于非线性可分情况, 此算法借助于核函数将原始问题映射到特征空间, 然后再按照线性可分情况求解. 由于支持向量往往分布在两类样本相邻的边界区域, 因此此方法可以较为精确地预选取支持向量. 在部分 UCI 标准数据集和 ORL 人脸数据库上的实验说明此算法较以往支持向量预选取算法可以更为快速准确地进行支持向量预选取.

关键词: 支持向量预选取; 随机中心; 距离排序; 边界样本集

中图分类号: TP3

文献标识码: A

文章编号: 1000-7180(2013)08-0036-04

A Random Center Distance Sorting-based Support Vector Pre-extracted Method

HU Zhi-jun, WANG Hong-bin, LI Rong

(Department of Computer Science, Xinzhou Teachers University, Xinzhou 034000, China)

Abstract: In this paper a random center distance sorting-based support vector pre-extracted method is proposed. For linear separable case, the method firstly randomly selected from each category a certain number of samples and calculated the mean of them as a random center of the category. Then the method calculated for each sample the distance between the sample and the random center of the other class. Finally select a certain number of original samples with small distances to compose boundary samples. For non-linear separable case, the method maps the original problem into feature space with kernel functions, and solves the problem like linear separable case. As support vectors often locate in the border area adjacent to the other type of samples, this method can pre-extract support vectors more exactly. Experiments on UCI standard data sets and ORL data set show that the proposed algorithm can pre-extract support vectors faster and more exactly than previous support vectors pre-extraction methods.

Key words: support vector pre-extraction methods; a random center; distance sorting; boundary samples set

1 引言

传统支持向量机算法需要求解核矩阵的逆, 因此时空复杂度分别为 $O(n^3)$ 和 $O(n^2)$, 这严重影响了支持向量机算法对海量数据的处理, 如何有效缩短支持向量机算法的训练时间以及减少训练时所需的存储空间就成了研究的热点. 由于支持向量机算

法的运行时间和存储空间仅仅与支持向量有关, 而且支持向量往往只占原始样本的很小一部分, 因此如果能预选取支持向量, 就可以很好地降低支持向量机算法训练的时空复杂度.

在预选取支持向量方法, 国内外学者做了很多研究. 主要可分为基于迭代策略的分解方法和基于几何关系的支持向量抽取方法^[1-3]. Shin 等人^[4]提

收稿日期: 2012-12-01; 修回日期: 2013-01-17

基金项目: 山西省自然科学基金项目(2009011018-4); 山西省高校科技开发项目(20101120)

出了一种基于 KNN 算法的支持向量预选取方法,该方法通过设定异类样本所占比例阈值来进行边界向量的选取,从而涵盖支持向量.然而该方法不能有效处理非线性可分的情况,而且通过 KNN 算法进行支持向量预选取会导致计算复杂度较高,为 $O(n^2)$.

本文提出了一种基于随机中心距离排序的支持向量预选取算法.

2 支持向量机理论简述

支持向量机是 Vapnik 基于统计学习理论提出的一种分类新方法,可以有效地处理高维、非线性和小样本数据,其主要依据结构风险最小化进行最优决策界面的确定.根据样本的可分性,支持向量机可以分为线性支持向量机和非线性支持向量机.

对于线性支持向量机,训练数据时线性可分的或者是近似线性可分的.此时支持向量机算法通过求解如下优化问题进行最优分类界面的确定.

$$\max W(a) = \sum_{i=1}^l a_i - \frac{1}{2} \sum_{i,j=1}^l a_i y_i a_j y_j (x_i \cdot x_j) \quad (1)$$

其中目标参数需要满足如下约束条件:

$$\sum_{i=1}^l a_i y_i = 0, a_i \in [0, C], i=1, \dots, l \quad (2)$$

式中, C 是惩罚因子, a_i 为拉格朗日乘子.

对于非线性支持向量机,由于训练数据不可分,因此支持向量机算法首先通过核函数隐式的将训练数据映射到高维特征空间,然后同样按照线性可分进行求解.此时的优化问题如下:

$$\text{Max} W(a) = \sum_{i=1}^l a_i - \frac{1}{2} \sum_{i,j=1}^l a_i y_i a_j y_j K(x_i \cdot x_j) \quad (3)$$

约束条件与线性可分情况下的约束条件一致.

当求出优化问题的解 a_i 之后,两种情况的决策超平面就可以分别表示如下:

$$f(x) = \text{sgn}(\sum_{i=1}^l a_i^* y_i (x_i \cdot x) + y_i - \sum_{i=1}^l a_i^* y_i (x_i \cdot x_j)) \quad (4)$$

$$f(x) = \text{sgn}(\sum_{i=1}^l a_i^* y_i K(x_i \cdot x) + y_j - \sum_{i=1}^l y_i a_i^* K(x_i, x_j)) \quad (5)$$

由于只有非零的拉格朗日乘子 a_i^* 才会影响最终的分类界面,而非零 a_i^* 对应的样本被称作支持向量,因此如果能对支持向量进行预选取就可以极大地降低支持向量机求解的时空复杂度.

3 基于随机中心距离排序的支持向量预选取算法

3.1 线性可分训练集

设样本集合为 $(x_i, y_i), i=1, \dots, l, x_i \in R^n, y_i \in \{-1, 1\}$. 记第一类样本个数为 l_1 , 第二类样本个数为 l_2 , 则 $l=l_1+l_2$. 对于线性可分训练集, 首先计算两类样本的随机中心, C_1 和 C_2 . 也就是从每一类样本中随机选取比例为 a 的样本求均值, 并以此均值作为随机中心. 计算公式如下:

$$C_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$$

$$C_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_j \quad (6)$$

其中

$$n_1 = l_1 a < l_1, n_2 = l_2 a < l_2 \quad (7)$$

然后分别计算每一个样本和另一类随机中心之间的距离. 不妨设第一类样本中的每一个和随机中心 C_2 之间的距离为 $d_i, i=1, \dots, l_1$, 则计算公式如下:

$$d_i = d(x_i, C_2) = \|x_i - C_2\|, i=1, \dots, l_1 \quad (8)$$

同理第二类样本中的每一个和中心 C_1 之间的距离 $d'_i, i=1, \dots, l_2$, 可以如下计算:

$$d'_i = d(x_i, C_1) = \|x_i - C_1\|, i=1, \dots, l_2 \quad (9)$$

最后设定抽取百分比阈值 b , 按照百分比阈值 b 从小到大抽取一定比例的小距离样本实现支持向量预选取. 为叙述方便, 称这些选择的样本组成的集合为边界样本集合. 如上所述, 最终边界样本集合的大小为 $m=l_1 b+l_2 b=l b$.

由于支持向量往往分布在靠近另一类样本的边界区域, 因此选择距离另一类样本随机中心较近的样本可以有效地涵盖支持向量.

3.2 非线性可分训练集

对于非线性可分训练集, 由于样本交叉在一起, 所以在原始空间很难寻找边界样本, 因此采用核函数将原始样本映射到高维特征空间, 使得原始样本在特征空间线性可分, 从而可以用类似于线性可分时的预选取方法.

对于非线性可分训练集, 最关键的是计算特征空间中样本与随机中心之间的距离. 下面给出计算方法. 设计算第一类随机中心随机抽取的样本集合为 $\{x_1, x_2, \dots, x_{n_1}\}$, 则特征空间中第一类随机中心计算公式如下:

$$C_1 = \sum_{i=1}^{n_1} \psi(x_i) \quad (10)$$

式中, n_1 是第一类随机抽取的样本的个数. 那么对于第二类中任意一个样本 x , 其特征空间对应的样本与第一类样本随机中心的距离可以如下计算:

$$\begin{aligned} d(\psi(x), C_1)^2 &= \left\| \psi(x) - \sum_{i=1}^{n_1} \psi(x_i) \right\|^2 \\ &= \left(\psi(x) - \sum_{i=1}^{n_1} \psi(x_i) \right) \cdot \left(\psi(x) - \sum_{i=1}^{n_1} \psi(x_i) \right) \\ &= \psi(x) \cdot \psi(x) - 2\psi(x) \cdot \sum_{i=1}^{n_1} \psi(x_i) + \\ &\quad \sum_{i=1}^{n_1} \psi(x_i) \cdot \sum_{i=1}^{n_1} \psi(x_i) \\ &= K(x, x) - 2 \sum_{i=1}^{n_1} K(x, x_i) + \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} K(x_i, x_j) \end{aligned} \quad (11)$$

同理可求出第一类样本点与第二类样本随机中心的距离. 这样就可以按照线性可分的方法进行支持向量预选.

3.3 算法复杂度分析

(1) 线性可分情况

① 计算随机中心的复杂度为 $O(la)$, 其中 l 为两类样本的总数目, a 为随机样本选取比例.

② 计算数据集中各个样本到另一类样本随机中心的距离的计算复杂度为 $O(l)$.

③ 从第一类样本中选取随机中心距离较小的前 $m_1 = l_1 b$ 个样本的计算复杂度为 $O(l_1 \log_2^{m_1})$, 从第二类样本中选取随机中心距离较小的前 $m_2 = l_2 b$ 个样本的计算复杂度为 $O(l_2 \log_2^{m_2})$. 该结论的主要证明过程如下.

证明: 因为需要对第一类样本中选取随机中心距离较小的前 m_1 个样本, 所以可以对任意 m_1 个样本的距离进行一次从小到大排序, 复杂度为 $O(m_1 \log_2^{m_1})$, 然后对剩余 $(l_1 - m_1)$ 个样本中的每一个样本的距离与排好顺序的 m_1 个距离进行比较, 若大于前 m_1 个样本的最大距离, 则不做操作, 若小于则进行折半查找, 最多以 $O(\log_2^{m_1})$ 的复杂度插入到前面的序列中, 然后将原来序列最后的样本剔除. 如此重复 $(l_1 - m_1)$ 次, 就可以获得前 m_1 个具有较小距离的样本. 因此复杂度为

$$O(m_1 \log_2^{m_1} + (l_1 - m_1) \log_2^{m_1}) = O(l_1 \log_2^{m_1}) \quad (12)$$

同理可知从第二类样本中选取随机中心距离较小的前 m_2 个样本的计算复杂度为 $O(l_2 \log_2^{m_2})$.

证毕.

如此就得到了线性可分情况, 本文支持向量预选算法的复杂度为

$$O(l_1 \log_2^{m_1} + l_2 \log_2^{m_2} + la + l) < O(l \log_2^l) \quad (13)$$

(2) 非线性可分情况:

① 由式(11)可知计算第二类样本中所有样本到另一类样本随机中心的距离总共需要做 $l_2 + l_2 n_1 + n_1^2$ 次核运算, 同样计算第一类中每一个样本到第二类样本随机中心的距离总共需要做 $l_1 + l_1 n_2 + n_2^2$ 次核运算. 然而由于本文仅仅需要比较每一类中各样本到另一类样本随机中心的距离大小, 由于式(11)中最后一项对于每一类是固定的, 因此可以不计算该项. 所以总共需要做的核运算的次数为

$$\begin{aligned} &l_2 + l_2 n_1 + l_1 + l_1 n_2 \\ &< l_2 + l_2 l_1 + l_1 + l_1 l_2 \\ &= l_1 + l_2 + 2l_1 l_2 \\ &< l^2 + l \end{aligned} \quad (14)$$

因此计算样本与随机中心的距离的复杂度将小于 $O(l^2)$.

② 如线性可分情况, 选取具有较小随机中心距离的样本的复杂度将小于 $O(l \log_2^l)$.

由于当 n 较大时, $l \log_2^l < l^2$, 所以非线性可分情况下, 本文支持向量预选算法的复杂度小于 $O(l^2)$.

4 实验结果与分析

4.1 实验一

本文首先利用 UCI 标准数据集对本文支持向量预选算法+支持向量机分类算法(SVC)与文献[3-4]中所提支持向量预选算法+ SVC 进行对比, 以验证本文算法能够更快更合理的确定支持向量. 所利用的 UCI 标准数据集分别是 vote, iris, vehicle, audiology, balance, breast, colic, credit, diabetes 和 glass. 其中 iris, vehicle, audiology, balance 和 glass 数据集属于多分类数据集, 其余的属于两分类数据集. 本实验均采用如下径向基函数作为 SVC 算法的核函数.

$$K(x, x_i) = \exp(-\|x - x_i\|^2 / \sigma^2) \quad (15)$$

对于其中的多分类问题则采用一对^[5]多分类方法. 表 1 中给出了三种算法在各个数据集上的训练时间和分类精度.

4.2 实验二

为了进一步验证本文算法在高维数据集上的有

效性,本文在 ORL 人脸数据库上进行第二个实验. ORL 人脸数据库包括 40 个人的人脸图像,其中每个人有 10 幅不同的图像. 这些图像在不同时间获取,具有不同的表情和面部细节,而且有一定程度的倾斜(不超过 20 度). 图 1 是其中的部分图片.

表 1 UCI 数据集上的实验结果

数据集	训练时间/s,分类精度/%		
	本文算法 +SVC	文献[3]算法 +SVC	文献[4]算 法+SVC
vote	2.1, 92.6	7.6, 92.1	8.2, 91.0
iris	1.2, 100	5.6, 100	6.5, 100
vehicle	5.5, 92.1	13.1, 90.1	10.2, 89.2
audiology	10.6, 89.9	12.1, 89.2	15.2, 88.2
balance	11.1, 95.6	13.2, 93.2	14.3, 95.1
breast	6.7, 93.4	8.8, 92.8	9.2, 92.0
colic	10.5, 98.1	12.3, 97.0	15.6, 96.1
credit	3.3, 92.9	6.7, 89.0	10.1, 91.0
diabetes	4.5, 89.8	8.8, 88.2	10.2, 86.9
glass	5.3, 94.5	9.5, 90.6	10.8, 92.7



图 1 ORL 人脸数据库部分样本

本实验首先对 ORL 人脸数据库的每幅图像均经下采样成为 64×64 的图像. 本实验取每个人的前 5 幅图像作为训练集,剩余的 5 幅图像作为测试集. 实验中仍选取高斯径向基函数作为核函数,实验采用 10 倍交叉验证方法^[6],结果如表 2 所示.

表 2 ORL 人脸数据库上的分类结果比较

算法	分类精度/%	训练时间/s
本文算法+SVC	93.7	59.2
文献[3]算法+SVC	87.9	100.3
文献[4]算法+SVC	88.1	121.3

4.3 实验结果分析

从表 1 和表 2 可以看出,经本文算法进行支持向量预选取然后进行标准支持向量机训练,所得到的分类精度普遍好于文献[3]和文献[4]对应的算法,并且所需要的训练时间也普遍少于文献[3]和文献[4]对应的算法. 这充分说明了本文算法可以更快地精确的进行支持向量预选取.

5 结束语

提出了一种基于随机中心距离排序的支持向量预选取算法. 该算法通过选取与另一类样本随机中心距离较小的样本来预选取支持向量,较以往支持向量预选取方法具有更高的效率和精度. 在部分 UCI 数据集和 ORL 人脸数据库上的对比试验充分说明了本文算法的有效性.

参考文献:

[1] Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers[C]//Proceedings Fifth Annual Workshop on Computational Learning Theory. New York, 1992:144—152.

[2] Osuna E, Freund R, Girosi F. An improved training algorithm for support vector machines [C]//IEEE Workshop on Neural Networks for Signal Processing. Amelia Island Florida, 1997:276—285.

[3] 李青,焦李成,周伟达. 基于向量投影的支持向量预选取[J]. 计算机学报, 2005, 28(2):145—152.

[4] Shin H, Cho S. Pattern selection for support vector classifiers[C]//Proc. of the 3rd International Conference on Intelligent Data Engineering and Automated Learning. Springer—Verlag, 2010: 469 - 474.

[5] Hsu C W, Lin C J. A comparison on methods for multi— class support vector machines [J]. IEEE Transactions on Neural Networks, 2001, 13(2): 415 —425.

[6] Kohavi R. A study of cross— validation and bootstrap for accuracy estimation and model selection[C]//Wermter S, Riloff E, Scheler G, eds. Proc. 14th Joint Int. Conf. Artificial Intelligence. San Mateo, CA: Morgan Kaufmann, 1995:1137—1145.

作者简介:

胡志军 男, (1974—), 硕士, 副教授. 研究方向为智能算法、人工神经网络.

王鸿斌 男, (1972—), 博士, 副教授. 研究为智能信息处理.

李荣 男, (1975—), 硕士, 副教授. 研究方向为人工智能及中文信息处理.