

基于核向量机的模式分类及其实验测试

郭瑞华,程国建

(西安石油大学 计算机学院,陕西 西安 710065)

摘 要:文中使用一种新的 SVM 变种——核向量机来对大样本数据集进行训练建模,进而求解模式分类问题. CVM 算法是将核函数转换为最小包围球问题进行求解,可以解决任何线性或非线性分类问题.测试结果表明,核向量机可以快速对大样本数据进行分类并能产生较少的支持向量.

关键词:核向量机;支持向量机;最小包围球;核函数

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2010)09-0190-03

Pattern Classification and Experiment Testing Based on Core Vector Machine

GUO Rui-hua, CHENG Guo-jian

(School of Computer, Xi'an Shiyou University, Xi'an 710065, China)

Abstract: To handle such kind of large datasets, we use a new kind of CVM variable — Core Vector Machine (CVM) for pattern classification. CVM algorithm formulates original kernel methods in SVM as a Minimum Enclosing Ball (MEB) problems in computational geometry and it can be used with any linear/nonlinear kernels. Experiment shows that the CVM is as accurate as existing SVM implementations, but it gets smaller support vectors and much faster than SVM for very larger datasets.

Key words: core vector machine (CVM); support vector machine (SVM); minimum enclosing ball (MEB); kernel functions

1 引言

在各种机器学习中核方法是最成功的一种算法,其中,支持向量机在解决分类、回归和模式识别等机器学习问题中表现出许多突出的优势.传统上, SVM 算法的本质是把核函数转换为标准二次规划 (Quadratic Programming, QP) 问题进行求解.而标准 QP 优化问题求解的时间复杂度为 $O(m^3)$,其空间复杂度至少是 $O(m^2)$ (其中 m 是训练样本数).显然,随着每类训练样本数的增多, QP 问题求解所需的内存空间快速增大,求解时间也快速增加.因此在实际应用中,这些方法只适用于小样本情况,当样本数目较大时,二次规划就成了 SVM 训练算法速

度慢的主要原因.

为了降低时间和空间复杂度,提出了许多针对大规模训练集的 SVMs 训练算法.最常用的就是分块算法或更复杂的分解算法.其中一个较受欢迎的算法是序贯最小优化算法 (Sequential Minimal Optimization, SMO),该算法基于 Osuna 的理论研究结论,其基本思想是把最初的一个大型 QP 问题分解为一系列最小规模的 QP 子优化问题进行求解,每次仅包含两个样本.其他分解算法的思想都类似.

Tsang 等人在 2005 年提出的一种大样本数据快速分类算法——核向量机 (Core Vector Machines, CVM)^[1]. CVM 结合了计算几何学和 SVM 训练算法^[2],是一种综合技术.它是把 SVM 训练的 QP 问

题转换为最小包围球 (Minimum Enclosing Ball, MEB) 问题, 然后使用了一种有效的 $(1 + \epsilon)$ 近似算法, 从而获得一个近似的 SVM 最优解. 文中使用 MIT 人脸数据库对 CVM 算法和 SVM 算法进行测试并进行对比分析. 实验表明, 与 SVM 相比, 核向量机可以快速对大样本数据进行分类并能产生较少的核向量, 并保证了分类的准确性.

2 核向量机的原理与算法

因为 CVM 本质是将核方法转化为 MEB 问题进行求解, 所以首先简要的介绍 MEB 问题, 然后介绍 CVM 算法.

2.1 MEB 的基本原理

考虑 m 个 D 维训练样本, 记作 $S = \{x_1, x_2, \dots, x_m\}$, 数据集 S 的最小包围球 $\text{MEB}(S)$ 定义为包含 S 中所有点的最小球^[3]. 当训练样本的维数 D 大于 30 时, 传统的寻找精确 MEB 的方法就会失效. 因此, 研究一种快速近似算法具有实际意义. 首先定义 $B(c, R)$ 是球心为 c , 半径为 R 的一个球体. 如果 $R \leq r_{\text{MEB}(S)}$ 且 $S \subset B(c, (1 + \epsilon)R)$, 则球 $B(c, (1 + \epsilon)R)$ 是 $\text{MEB}(S)$ 的 $(1 + \epsilon)$ 逼近, 其中 $\epsilon > 0$. 从许多拟合问题中发现, 如果把对整个数据集 S 的求解转化成对 S 的一个子集 Q 求解, 会得到一个精确有效的近似解, 其中 Q 被称为核集. 更正式的定义是: 对于 $Q \subseteq S$, 如果 $S \subset B(c, (1 + \epsilon)R)$, 其中 $B(c, R) = \text{MEB}(Q)$, 则 Q 是 S 的一个核集. 最小包围球如图 1 所示.

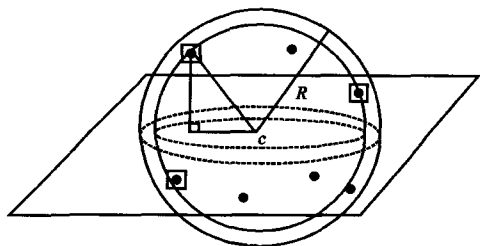


图1 最小包围球

在图 1 中, 内圆是那些用正方形标记点的最小包围球, 而外圆则是内圆的 $(1 + \epsilon)$ 扩展, 其中外圆包含了所有的点, 因此被正方形标记的点的集合是核集^[4-5].

$(1 + \epsilon)$ 近似 MEB 算法是由 Badoiu 和 Clarkson 在 2002 年提出的, 使用了简单的迭代策略: 即在 t 次迭代后得到的 MEB 球为 $B(c_t, r_t)$. 之后每次迭代都是把在 $B(c_t, (1 + \epsilon)r_t)$ 之外的一个最远的点包括进来. 直到所有的点都包括在 $B(c_t, (1 + \epsilon)r_t)$

中, 则迭代结束. 尽管这个算法很简单, 但是它的一个突出特点是迭代的次数以及最终核集的大小仅依赖 ϵ , 而与 D 和 m 无关.

2.2 核向量机的基本原理

给定一组点集 $S = \{x_1, \dots, x_m\}$, 其中每一个 $x_i \in R^d$, 核向量机通过特征映射 ϕ 将样本投影到一个高维特征空间并建立最小包围球 $\text{MEB}(S)$, 其中球心和半径分别为 c 和 r . 在特征空间中寻求 $\text{MEB}(S)$ 问题等价于如下优化问题^[6]:

$$\begin{aligned} \min \quad & r^2 \\ \text{s. t.} \quad & \|\phi(x_i) - c\| \leq r^2, \quad \forall i = 1, \dots, M \end{aligned} \quad (1)$$

利用 Lagrange 优化方法可以把 MEB 问题转换为较简单的对偶问题, 并描述如下:

$$\begin{aligned} \max_a \quad & \sum_{i,j=1}^M a_i a_j K(x_i, x_j) - \sum_{i=1}^M a_i K(x_i, x_i) \\ \text{s. t.} \quad & \sum_{i=1}^M a_i = 1, \quad a_i \geq 0, \quad \forall i = 1, \dots, M \end{aligned} \quad (2)$$

其中 a_i 是 Lagrange 乘子, $K(x_i, x_j)$ 是核矩阵. 如果把 $K(x_i, x_j)$ 记为常数 k , 即:

$$K(x_i, x_j) = k \quad (3)$$

则有 $\sum_{i=1}^M a_i K(x_i, x_j) = k$. 如果把式(2)中的该项去掉, 就会得到较简单的 $\text{MEB}(S)$ 优化问题:

$$\begin{aligned} \max_a \quad & \sum_{i,j=1}^M a_i a_j K(x_i, x_j) \\ \text{s. t.} \quad & \sum_{i=1}^M a_i = 1, \quad a_i \geq 0, \quad \forall i = 1, \dots, M \end{aligned} \quad (4)$$

无论在什么时候, 只要 k 满足式(3), 则任何具有式(3)形式的 QP 问题都可以认为是 MEB 问题. 这里 CVM 选择的核函数为:

$$\tilde{K}(x_i, x_j) = y_i y_j K(x_i, x_j) + y_i y_j + \frac{\delta_{ij}}{C} \quad (5)$$

因为 k 满足式(3), $\tilde{K}(x_i, x_j) = k + 1 + \frac{1}{C}$ 是一个常数, 所以式(4)是一个 MEB 问题. 然后 CVM 使用 $(1 + \epsilon)$ 近似迭代算法求解 MEB 问题. 算法具体步骤如下:

(1) 初始化 $S_0 = \{\phi(Z_0)\}$, $c_0 = \phi(Z_0)$ 和 $R_0 = 0$.

(2) 如果没有任何点 $\phi(z)$ 落在球 $B(c_t, (1 + \epsilon)r_t)$ 外面, 则结束. 否则, 选择离球心最远的点 $\phi(z_t)$, 使集 $S_{t+1} = S_t \cup \{\phi(z_t)\}$.

(3) 寻找新的 $\text{MEB}(S_{t+1})$, 其中 $c_{t+1} = c_{\text{MEB}(S_{t+1})}$ 和 $R_{t+1} = R_{\text{MEB}(S_{t+1})}$.

(4) 使 t 增 1, 然后转到步骤(2).

其中定义第 t 次迭代时, MEB 球的核集、球心和半径分别为 S_t, c_t 和 R_t . 而球 B 的球心和半径分别为 c_B 和 r_B .

3 实验与分析

CVM 是根据 LIBSVM 改编并通过 SMO 算法来解决 MEB 问题而实现的. 正如 LIBSVM, CVM 使用缓存把所有训练样本存储在主存里并使用概率加速方法进行计算. 为验证文中算法的性能, 采用下面四种数据集作为测试数据. 这些测试数据集的主要特征描述如表 1 所示.

表 1 测试数据集描述

Data Sets	# class	dim	# size
letter	26	16	15 000
web	2	300	49 749
Forest cover type	2	54	522 910
Extended usps	2	676	266 079

分类器的预测准确率度量采用 5 次交叉验证方

法. 在这种方法中每条数据既做了训练数据, 又做了测试数据, 因此, 它能较全面的测试算法的性能.

实验测试的硬件平台是 PC 机 (P4 1.79GHz, 512MB RAM, Windows XP).

该实验分别使用 CVM 和 SVM 两种分类方法进行对比测试. 这两种算法都采用高斯核函数 $k(x, y) = \exp(-\|x - y\|^2/\beta)$, 其中 $\beta = \frac{1}{m^2} \sum_{i,j=1}^m \|x_i - x_j\|^2$. 而惩罚因子 C 和核参数 σ 是通过训练数据的交叉验证选取的, 实验发现, 除了 Forest cover type 数据集, 把惩罚因子 $C = 100$, 核参数 $\sigma = 0.097$ 作为模型的最终参数可以使其余数据集达到比较好的效果. 表 2 列出了这两种方法的平均分类正确率与标准方差.

可以看出, 对于较大样本数据集 (letter 和 web), SVM 算法和 CVM 算法的分类精确度以及产生的支持向量大致相同; 而对于大样本数据集 Extended usps 和 Forest cover type, 与 SVM 相比, CVM 则产生了较少的支持向量, 但是其精度还是与 SVM 相当.

表 2 分类正确率比较

Data sets	SVM		CVM	
	Accuracy %	SV	Accuracy %	CV
letter	97.350 ± 2.57	6238	97.425 ± 2.38	6806
web	99.452 ± 0.021	3987	99.445 ± 0.022	4314
Forest cover type	86.432% ± 0.017	246834	86.543% ± 0.02	7328
Extended usps	99.536 ± 0.019	2043	99.556 ± 0.017	1065

CVM 算法的最大优势是能够快速的处理大样本数据, 时间复杂度较低. 所以本实验也对算法的运行时间进行比较, 在所有测试集上的运行时间如表 3 所示.

表 3 运行时间比较 (s)

Data set	letter	web	Forest cover type	Extended usps
CVM	20.094	113.094	1086.563	203.093
SVM	20.625	189.141	102009.156	2052.55

可以看出, CVM 在这四种数据集上的运行时间都低于 SVM, 尤其对于大样本数据集 (Forest cover type 和 Extended usps), CVM 的运行时间远远低于 SVM 的运行时间. 原因在于 CVM 把 SVM 训练的 QP 问题转换为 MEB 问题, 并与计算几何学相结合. CVM 能够为每类数据构建一个最小包围球, 它可以有效减少支持向量数目并把目标函数与奇异点进行分离, 减少对一些无关数据不必要的训练时间,

从而节约运行时间.

4 结束语

核向量机是一种新的模式识别方法, 它通过将核方法转换为等价的 MEB 问题, 并与计算几何学相结合, 从而解决了处理大样本所需要的大量内存空间的问题. 实验数据表明, 与 SVM 相比, CVM 算法保证了分类准确率, 并且可以有效地节约大量的内存时间. 与理论上证明其时间复杂度仅为 $O(m)$, 而空间复杂度与 m 无关. 计算效率高是 CVM 成为研究热点的主要原因之一.

参考文献:

- [1] 李瑜, 郑敏娟, 程国建. 神经网络方法拟合核磁共振测井曲线[J]. 微电子学与计算机, 2009, 26(2): 19-21.

(下转第 196 页)

参考文献:

- [1] 汤晖, 许伦辉, 汤青波. 退火遗传算法在交通图像分割中的应用研究[M]. 微电子学与计算机, 2008, 25(8): 23 - 25.
- [2] Goldberg D E. Genetic algorithms in search, optimization and machine learning [M]. Boston: Addison - Wesley publishing, 1989.
- [3] 张海建, 陈向东, 王世海. 一种遗传算法优化的图像模糊边缘检测方法[M]. 微电子学与计算机, 2008, 25(4): 105 - 107.
- [4] 严勇. 一种求解航线优化问题的改进蚁群遗传算法[J]. 微电子学与计算机, 2009, 26(2): 115 - 117.
- [5] Hwang S F, He R S. Improving real - parameter genetic algorithm with simulated annealing for engineering problem [J]. Advances in Engineering Software, 2006(37): 406 - 418.
- [6] Zhang J, Chung H S H, Lo W L. Clustering - based adaptive crossover and mutation probabilities for genetic algorithms[J]. IEEE Transactions on Evolutionary Computation, 2006, 11(3): 326 - 335.
- [7] Hinterding R, Michalewicz Z, Eiben A. Adaptation in evolutionary computation: a survey [C]// Proceedings of IEEE International Conference on Evolutionary Computation. Piscataway. NJ., 1997: 65 - 69.

作者简介:

黄江波 男, (1975 -), 硕士, 讲师. 研究方向为人工智能与电子技术检测.

(上接第 192 页)

- [2] Ivor W Tsang, James T Kwok, Pak - Ming Cheung. Core vector machines: Fast SVM training on very large data sets [J]. Journal of Machine Learning Research, 2005(6): 363 - 392.
- [3] Lessmann, Stefan Li, Ning Voss. A case study of core vector machines in corporate data mining[C]// Proceedings of the 41st Annual Hawaii International Conference on System Sciences. Hawaii, 2008: 78 - 78.
- [4] Ivor W Tsang, Andras Kocsor James T Kwok. Simpler core vector machines with enclosing balls[C]// Proceedings of the Twenty - Fourth International Conference on Machine Learning (ICML). Corvallis, Oregon, USA, 2007.
- [5] Asharaf S, Narasimha Murty M, Shevade S K. Multiclass core vector machine[C]// Proceedings of the 24th international conference on Machine learning. Corvallis, Oregon, 2007: 41 - 48.
- [6] Ivor W Tsang, James T Kwok, Jacek M Zurada. Generalized core vector machines[J]. IEEE Transactions on Neural Networks, 2006, 17(5): 1126 - 1140.

作者简介:

郭瑞华 女, (1983 -), 硕士. 研究方向为神经网络、机器学习、软计算方法.

程国建 男, (1964 -), 博士, 教授. 研究方向为神经网络、进化计算、软计算方法在油藏描述中的应用.