文章编号: 1671- 251X(2010)12- 0047- 04

一种半监督支持向量机优化方法

王永[1], 程灿[2], 戴明军[2], 孙永[2]

(1. 华晋焦煤有限责任公司, 山西 河津 043300; 2. 中国矿业大学信电学院, 江苏 徐州 221008)

摘要: 针对半监督支持向量机在采用间隔最大化思想对有标签样本和无标签样本进行分类时面临的非凸优化问题, 提出了一种采用分布估计算法进行半监督支持向量机优化的方法 EDA_S^{3}VM。该方法把无标签样本的标签作为需要优化的参数, 从而得到一个在标准支持向量机上的组合优化问题, 利用分布估计算法通过概率模型的学习和采样来对问题进行求解。在人工数据集和公共数据集上的实验结果表明, EDA_S^{3}VM 与其它一些半监督支持向量机算法相比有更高的分类准确率。

关键词: 半监督学习; 支持向量机; 分布估计算法; 组合优化

中图分类号: **TP181** 文献标识码: **A**

An Optimized Method for Semi-supervised Support Vector Machines

WANG Yong[1], CHENG Can[2], DAI Ming-jun[2], SUN Yong[2]

(1. Huajin Coking Coal Co., Ltd., Hejin 043300, China.
2. School of Information and Electrical Engineering of CUMT., Xuzhou 221008, China)

Abstract: In view of problem of non-convex optimization problem that semi-supervised support vector machines use margin maximization principle to classify labeled and unlabeled samples, a method EDA_S^{3}VM was proposed which using estimation of distribution algorithm to optimize semi-supervised support vector machines. Labels of unlabeled samples are taken as optimized parameters to obtain a combinatorial optimization problem on standard support vector machines, which can be solved by estimation of distribution algorithm through learning and sampling of probability model. The experiment results of artificial and UCI datasets showed that EDA_S^{3}VM has better classification accuracy than other methods of semi-supervised support vector machines.

Key words: semi-supervised learning, support vector machine, estimation of distribution algorithm, combinatorial optimization

0 引言

目前, 半监督学习在机器学习领域发展迅速, 主要是由于无标签样本的获得较容易并且经济。半监督学习就是通过少量的有标签样本和大量的无标签样本来学习分类机, 从而获得更好的推广能力。半监督支持向量机是采用支持向量机来解决半监督问题的方法, 已经有大量的学者从事这方面的研究, 并取得了很好的效果。首次相近的研究是 Vapnik[1] 提出的直推式支持向量机。Joachims[2] 首次编码实现了支持向量机软件包 svmlight(其中包括直推式支持向量机)后, 大量解决半监督支持向量机的非凸问题的研究随后出现。相关的研究包括局部组合搜索[2]、梯度下降[3]、连续优化技术[4]、凸凹过程[5-6]、半正定编程[7-8]、不可微方法[9]、决定退火[10]、分枝界定法[11-12]。与支持向量机不同的是半监督支持向量机的原始问题是非凸的, 上面所提到的方法都是为解决这个非凸问题提出的。

分布估计算法是进化计算领域新兴的一类随机优化算法, 是当前国际进化计算领域的研究热点。

收稿日期: 2010- 10- 31

作者简介: 王永(1962-), 男, 山西大同人, 1985 年毕业于大同煤炭工业学校机电专业, 2000 年毕业于太原理工大学工业企业管理专业, 现为华晋焦煤有限责任公司机电副总工程师兼王家岭煤矿矿长, 主要从事煤矿管理及生产经营管理工作。E-mail: zjh4062@sina.com

它是遗传算法和统计学习的结合, 通过统计学习的手段建立解空间内个体分布的概率模型, 然后对概率模型随机采样产生新的群体, 如此反复进行, 实现群体的进化[13]。

因为半监督支持向量机的原始问题最终可归结为一个组合优化的问题, 而分布估计算法是可以解决组合优化问题的一种方法。本文的目的就是采用分布估计算法来解决半监督支持向量机的组合优化问题, 基于该原理, 提出了一种基于分布估计算法的半监督支持向量机优化方法 EDA_S^3VM, 给出了理论分析和实验结果。实验结果表明, EDA_S^3VM 与其它一些半监督支持向量机算法相比有更高的分类准确率。

1 半监督支持向量机

半监督支持向量机是采用间隔最大化思想对少量的有标签样本和大量无标签样本进行分类。与支持向量机不同, 除了确定 ω 和 b (ω 为垂直于分类超平面的向量, b 为一个分类阈值) 外, 半监督支持向量机把无标签样本的标签作为需要优化的参数。从而得到一个在标准支持向量机上的组合优化问题。下面给出了半监督支持向量机的工作原理, 这里只考虑二分类问题。

训练样本包括 l 个有标签样本$\{(x_i, y_i)\}_{i=1}^{l}$, $y_i = \pm 1$, u 个无标签样本$\{x_i\}_{i=l+1}^{n}$, 其中 $n = l + u$。在线性情况下, 目标函数为

$$\min_{\omega, b, \boldsymbol{y}_u, \xi_i} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i^p + C^*\sum_{i=l+1}^{n}\xi_i^p \quad (1)$$

$$\text{s.t. } y_i(\omega x_i + b) \geqslant 1 - \xi_i$$

式中: $\boldsymbol{y}_u = [y_{l+1} \cdots y_n]^{\mathrm{T}}$, 为无标签样本的标签; C、C^* 分别为有标签样本和无标签样本上的征罚参数; ξ_i 为样本 x_i 对应的松弛变量, $\xi_i \geqslant 0$, $1 \leqslant i \leqslant n$; p 为 ξ_i 上的指数级。

由式(1) 可知

$$\xi_i = L[1 - y_i(\omega x_i + b)] \quad (2)$$

其中:

$$L(\square) = \begin{cases} \square, & \square \geqslant 0 \\ 0, & \square < 0 \end{cases} \quad (3)$$

则问题转化为

$$\min_{\omega, b, \boldsymbol{y}_u} \frac{1}{2}\omega^2 + C\sum_{i=1}^{l}L[1 - y_i(\omega x_i + b)]^p + C^*\sum_{i=l+1}^{n}L[1 - y_i(\omega x_i + b)]^p \quad (4)$$

取 $p = 2$, 并且令 $C = C^*$, 问题进一步转化为

$$\min_{\omega, b, \boldsymbol{y}_u} \frac{1}{2}\omega^2 + C\sum_{i=1}^{n}L[1 - y_i(\omega x_i + b)]^2 \quad (5)$$

令函数:

$$I(\omega, b, \boldsymbol{y}_u) = \frac{1}{2}\omega^2 + C\sum_{i=1}^{n}L[1 - y_i(\omega x_i + b)]^2 \quad (6)$$

原问题变为最小化函数 I, 注意到 I 有 3 个变量 ω、b 和 $\boldsymbol{y}_u$。

解决思路有 2 种:

(1) 连续优化

将原问题中的 $\boldsymbol{y}_u$ 用 $\mathrm{sgn}(\omega x_i + b)$ 来代替, 这样原问题就转化为一个在 ω 和 b 上的连续优化问题。采用这种方法的有参考文献[3] ~ [6]。

(2) 组合优化

对于给定的一个 $\boldsymbol{y}_u$, 在 ω 和 b 上的优化就是一个标准的支持向量机。定义

$$J(\boldsymbol{y}_u) = \min_{\omega, b} I(\omega, b, \boldsymbol{y}_u) \quad (7)$$

原问题变为在一组有限空间中求最小化 J 的组合优化问题, 对于每个 J 的求解都是一个标准支持向量机。采用这个思路的方法有参考文献[2]、[7] ~ [10]。本文采用的方法就是基于这一思路。

为了防止不平衡结果的发生, 需要考虑平衡约束:

$$\frac{1}{u}\sum_{i=l+1}^{n}\max(y_i, 0) = r \quad (8)$$

式中: r 为有标签样本中正标签所占的比例。

2 分布估计算法

分布估计算法采用类似遗传算法中的种群进化模式, 通过概率模型的学习和采样来对问题进行求解。它通过一个概率模型描述候选解在空间的分布, 采用统计学习手段从群体宏观的角度建立一个描述解分布的概率模型, 然后对概率模型随机采样产生新的种群, 如此反复进行, 实现种群的进化, 直到终止条件[13]。

分布估计算法的基本步骤如下:

(1) 基于图的群体初始化

考虑到样本的分布, 在进行群体初始化时采用聚类假设, 即距离近的样本具有相同的类别标签。这样给每个样本赋一个标签, 每个基因位置的概率值是通过计算它周围相应标签的比例而产生的。例如通过 Dijkastra 求得所有样本的距离矩阵 $\boldsymbol{E}$, 得到概率向量 $p = (p_1, p_2, \cdots, p_u)$, 根据该概率模型产生初始群体。

(2) 选择优势群体，更新概率模型

通过适应值函数 $J(\boldsymbol{y}_u)$ 计算各个个体的适应值，如果符合终止条件，输出结果，否则继续执行。适应值函数是采用间隔最大化思想的函数，因此，选择适应值最大的 N 个个体组成优势群体 D_t^S（t 表示群体进化到了第 t 代；S 为抽样数），由 D_t^S 估计联合概率分布，更新概率模型 p。

(3) 随机采样

根据概率模型 p 采样 M 次，得到新一代群体，返回步骤(2)继续执行。

3　半监督支持向量机优化方法 EDA- S^3VM

式(7)为目标函数，$\boldsymbol{y}_u \in \{-1, 1\}^u$，描述解空间的概率模型用简单的概率向量 $p = (p_1, p_2, \cdots, p_u)$ 表示，p 表示群体的概率分布，$p_i \in [0, 1]$，表示基因位置 i 取 1 的概率，$1 - p_i$ 表示基因位置 i 取 0 的概率。

基于分布估计算法的半监督支持向量机优化方法 EDA- S^3VM 的伪代码如图 1 所示。

输入：l 个有标签的样本 $\{(x_i, y_i)\}_{i=1}^{l}$，$y_i = \pm 1$，u 个无标签的样本 $\{x_i\}_{i=l+1}^{n}$，其中 $n = l + u$
输出：ω、b 和 $\boldsymbol{y}_u$
1　$t = 0$
2　do {
3　if $t = 0$
4　初始化群体 D_0
5　利用标准支持向量机求解每个个体的适应值
6　else
7　通过概率模型抽样得到 D_{Sampled}
8　利用标准支持向量机求解每个个体的适应值
9　利用更新方法通过 D_{t-1}、D_{Sampled} 和 D_{t-1}^S 生成新群体 D_t
10　根据选择方法选择集合 D_t^S
11　根据学习方法计算 D_t^S 的概率模型
12　$t = t + 1$
13　} until 终止标准成立

图 1　基于分布估计算法的半监督支持向量机优化方法 EDA- S^3VM 的伪代码

4　实验结果及分析

4.1　数据集

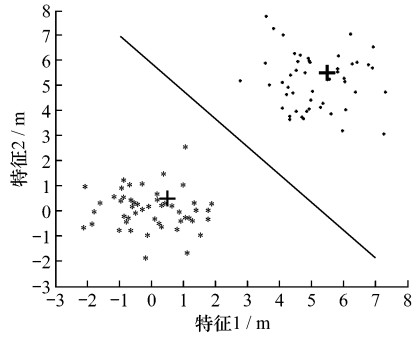
实验的目的是检验采用分布估计算法进行半监督支持向量机优化后在不同数据集上的分类效果。数据集包括 2 个人工数据集和 3 个公共数据集。其中人工数据集包括一个线性可分的点集(下面用 Point 表示)和一个“Two moons”数据集，公共数据集包括 g50c、Ciol20 和 Uspst。表 1 给出了实验数据集的特征描述，其中，c 表示样本类别的个数，d 表示样本的特征数，l 表示有标签的样本数，n 表示所有样本总数。

表 1　实验数据集的特征描述

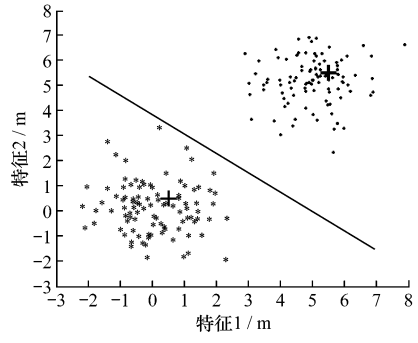
数据集	线性/非线性	c	d	l	n
Point	线性	2	2	2	102/202
Two moons	非线性	2	2	2	100
g50c	非线性	2	50	50	550
Ciol20	非线性	20	1 024	40	1 440
Uspst	非线性	10	256	50	2 007

4.2　实验结果及分析

对于 Point 数据集，在 2 个样本集上进行，其中一个数据集包括 2 个有标签样本(正负类各一个)、100 个无标签样本，另外一个数据集包括 2 个有标签样本(正负类各一个)、200 个无标签样本，结果如图 2 所示，其中右上方的“+”表示正类有标签样本，左下方的“+”表示负类有标签样本，右上方的“·”表示正类无标签样本，右下方的“*”表示负类无标签样本。图 2(a)中，EDA- S^3VM 找到了 2 类样本的最优分类面，而图 2(b)中，EDA- S^3VM 的分类效果较差，并未找到最优解，原因是分布估计算法采用的概率模型的学习和采样方法可能导致算法陷入局部最优。



(a) 100 个无标签样本



(b) 200 个无标签样本

图 2　EDA_S^3VM 的分类结果

然后, 通过实验来验证 EDA – S^3VM 算法在 4 种真实数据集上对无标签样本的分类错误率, 并与 cS^3VM、∇S^3VM 和 S^3VM^{light} 算法进行比较。由于这些算法的效果都会受参数 σ 与 C 取值的影响, 对每个数据集, 每一种算法在同样参数设置下采用交叉验证的方式取实验结果的平均值。在实验中进行如下设置: 粒子的种群规模对算法的运行速度有直接影响, 一般可取 10~ 40, 本实验取 20; 算法的终止条件为达到最大迭代次数或连续 10 次解不发生变化。表 2 给出了不同算法的分类错误率。

表 2 不同算法的分类错误率

数据集	cS^3VM	∇S^3VM	S^3VM^{light}	EDA_S^3VM
Two moons	45.7	59.3	66.2	35.2
g50c	50.0	55.4	60.1	45.4
Ciol20	60.6	60.6	55.3	41.2
Uspst	58.5	62.2	58.0	38.5

从表 2 可看出, EDA_S^3VM 与其它几种算法相比, 在测试样本上的分类错误率都有所降低。

5 结语

针对分布估计算法可解决半监督支持向量机的组合优化问题, 提出了一种基于分布估计算法的半监督支持向量优化方法 EDA_S^3VM。在人工数据集和公共数据集上的实验结果表明, EDA_S^3VM 与其它一些半监督支持向量机算法相比有更高的分类准确率。该方法可以推广到多分类问题与非线性问题, 对于多分类的问题采用一对多的策略变为二分类问题, 对于非线性的问题采用核方法来解决。由于分布估计算法在解决非线性、变量耦合的优化问题上的优势, 其在非线性支持向量机优化上的应用能得到较好的效果, 但由于进化算法的迭代过程导致了算法计算效率的下降, 如何加快算法的收敛速度是进一步研究的方向。

参考文献:

[1] VAPNIK V, STERIN A. On Structural Risk Minimization or Overall Risk in a Problem of Pattern Recognition[J]. Automation and Remote Control, 1977, 10(3): 1495-1503.

[2] JOACHIMS T. Transductive Inference for Text Classification Using Support Vector Machines[C]// Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, 1999.

[3] CHAPELLE O, ZIEN A. Semi-supervised Classification by Low Density Separation [C]// Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, Barbados, 2005.

[4] CHAPELLE O, CHI M, ZIEN A. A Continuation Method for Semi-Supervised SVMs [C]// International Conference on Machine Learning, Pennsylvania, 2006.

[5] FUNG G, MANGASARIAN O. Semi-supervised Support Vector Machines for Unlabeled Data Classification [J]. Optimization Methods and Software, 2001(15): 29-44.

[6] COLLOBERT R, SINZ F, WESTON J, et al. Large Scale Transductive SVMs[J]. Journal of Machine Learning Research, 2006(7): 1687-1712.

[7] DE BIE T, CRISTIANINI N. Semi-supervised Learning Using Semi-definite Programming[M]// CHAPELLE O, SCHOELKOPF B, ZIEN A. Semi-supervised Learning. Cambridge: MIT Press, 2006.

[8] XU L, NEUFELD J, LARSON B, et al. Maximum Margin Clustering [C]// Advances in Neural Information Processing Systems, Vancouver, 2004.

[9] ASTORINO A, FUDULI A. Nonsmooth Optimization Techniques for Semi-supervised Classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(12): 2135-2142.

[10] SINDHWANI V, KEERTHI S, CHAPELLE O. Deterministic Annealing for Semi-supervised Kernel Machines[C]// International Conference on Machine Learning, Pennsylvania, 2006.

[11] BENNETT K, DEMIRIZ A. Semi-supervised Support Vector Machines [C]// Advances in Neural Information Processing Systems, Colorado, 1998.

[12] CHAPELLE O, SINDHWANI V, KEERTHI S. Branch and Bound for Semi-supervised Support Vector Machines [C]// Advances in Neural Information Processing Systems, Vancouver, 2006.

[13] 周树德, 孙增圻. 分布估计算法综述[J]. 自动化学报, 2007, 33(2): 113-124.