

双目标函数支持向量机在情感分析中的应用

刘春雨¹, 朱倩男², 郭满才¹, 周晓飞²

(1 西北农林科技大学 理学院, 陕西 杨陵 712100; 2 中科院信息工程研究所, 北京 100093)

摘 要: 针对单目标函数支持向量机在情感分析中应用的不足, 采取两种双目标函数支持向量机 NPSVM 和 TWSVM 进行情感分类应用研究. 实验表明双目标函数支持向量机在应用于情感分析中, 其精度和稳定性具有明显优势.

关键词: 情感分析; 双目标函数; 支持向量机; 文本分类

中图分类号: TP301

文献标识码: A

文章编号: 1000—7180(2016)01—0127—04

Application of Twin Objective Function SVM in Sentiment Analysis

LIU Chun-yu¹, ZHU Qian-nan², GUO Man-cai¹, ZHOU Xiao-fei²

(1 Faculty of Science, Northwest A&F University, Yangling 712100, China;

2 Institute of Information Engineering, Chinese Academy Sciences, Beijing 10089, China)

Abstract: The paper studies sentiment analysis by applying the twin objective function SVM, including NPSVM and TWSVM. From the experiments, we concluded that twin objective function SVM are prior to single objective function SVM in accuracy and stability.

Key words: sentiment analysis; twin objective functions; SVM; text classification

1 引言

随着互联网和 web2.0 的蓬勃发展, 情感分析技术在用户评论分析和决策、舆情监控、信息预测等领域有很大的应用.

情感信息分类是情感分析技术研究主要任务之一, 即对主观信息的褒贬二元数据进行分类. 目前基于监督学习的情感分析是基于机器学习研究情感的热点. 文献[2]中首次基于文本的 N 元语法和词类等特征分别使用朴素贝叶斯、最大熵和支持向量机对文本进行情感正向和负向的倾向性分类. 目前, 基于监督学习的情感分析仍然是主流^[3,6]. 但是, 单一的目标函数旨在构造数据的决策面, 却没有很好地拟合数据, 并且运行速度慢, 适用数据小.

双目标函数的支持向量机是目前支持向量机领

域研究的热点. 该方法的目标函数有两个, 目的是产生两个分割超平面, 其优势是能更多地考虑数据的特征, 便于并行计算, 使得分类结果更准确, 运行速度更快, 对数据的适用范围更广. 本文将两种一般代表性的双目标函数的支持向量机应用到情感分析中. 这两种方法分别是 R. K. Jayadeva^[4]提出的孪生支持向量机(TWSVM)和 Tian^[5]等人提出的非平行支持向量机(NPSVM). 实验中, 将经典的支持向量机 libSVM、SMO 与本文应用方法进行比较. 双目标函数支持向量机具有精度高, 运行稳定等优势.

2 支持向量机

支持向量机是 Vapnik 在 1996 年针对线性不可分的两类问题的最优分类形式, 而提出来的分类器, 目标是为了找到一个最优分割超平面. 即通过

构造分类超平面使两类无错误地分开,并使两类的分类间隙最大.

对于给定训练数据集 $T = \{(x_i, y_i) \mid i = 1, \dots, l\}$, 其中 $x_i \in \mathcal{R}^n, y_i \in Y = \{+1, -1\}, i = 1, \dots, l$. 线性 SVM 解决以下原 QPP:

$$\begin{aligned} \min_{w, b, \xi} \quad & \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^l \xi_i \\ \text{s. t. } & y_i((w \cdot x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \dots, l \end{aligned} \quad (1)$$

式中, C 是惩罚参数, ξ_i 是收缩变量. 式(1)的对偶形式由式(3)表示:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{j=1}^l \alpha_j - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j (x_i \cdot x_j) \alpha_i \alpha_j \\ \text{s. t. } & \sum_{i=1}^l y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, i = 1, \dots, l \end{aligned} \quad (2)$$

式中, $\alpha \in \mathcal{R}^l$ 是 Lagrangian 因子. 根据优化问题(3)的解 α^* , 可求得最优分类超平面的参数 w 和 b :

$$w = \sum_{i=1}^l \alpha_i^* y_i x_i, \quad b = \frac{1}{N_{sv}} (y_j - \sum_{i=1}^{N_{sv}} \alpha_i^* y_i (x_i \cdot x_j)), \quad (3)$$

式中, N_{sv} 代表支持向量个数. 最终的判定函数为

$$f(x) = \text{sgn}((w \cdot x) + b) \quad (4)$$

LibSVM 和 SMO 是目前解决 SVM 的有效算法. LibSVM 是台湾大学林智仁等开发设计的一种简单快速、有效且易于使用的 SVM 的模式识别和回归的软件包. 其涉及参数调节较少, 并提供了交互检验的功能. SMO 是支持向量机的一种快速有效的算法. 它把大的凸规划问题转化为一个个小的凸规划问题, 且能训练较大的数据集. 本文对 SVM 的实现主要采用这两种方法.

3 双目标函数支持向量机

单目标函数支持向量机用于情感分析时存在以下的问题: 单一的目标函数旨在构造数据的决策面, 没有很好地拟合不同情感类别数据的分布. 双目标函数支持向量机在可以有效逼近数据分布的同时, 实现数据的最优分类.

3.1 TWSVM

孪生支持向量机(TWSVM)是 R. K. Jayadeva, 于 2007 年在支持向量机的基础收提出来的, 其目标是对二元分类找到两个超平面, 每个超平面是接近其中一类并尽可能远离另外一类.

考虑给定的训练集的二分类问题:

$$T = \{(x_1, +1), \dots, (x_p, +1), (x_{p+1}, -1), \dots, (x_{p+q}, -1)\} \quad (5)$$

找到两个超平面 $(w_+ \cdot x) + b_+ = 0$ 和 $(w_- \cdot x) + b_- = 0$, 这两个超平面可以被以下两个 QPPs 解决:

$$\begin{aligned} \min_{w_+, b_+, \xi_-} \quad & \sum_{i=1}^p ((w_+ \cdot x_i) + b_+)^2 + d_1 \sum_{j=p+1}^{p+q} \xi_j \\ \text{s. t. } & (w_+ \cdot x_i) + b_+ \leq -1 + \xi_j \\ & \xi_j \geq 0, j = p+1, \dots, p+q \end{aligned} \quad (6)$$

$$\begin{aligned} \min_{w_+, b_+, \xi_{+i=p+1}} \quad & \sum_{j=p+1}^{p+q} ((w_- \cdot x_j) + b_-)^2 + d_2 \sum_{j=1}^p \xi_j \\ \text{s. t. } & (w_- \cdot x_j) + b_- \leq 1 - \xi_j \\ & \xi_j \geq 0, j = 1, \dots, p \end{aligned} \quad (7)$$

式中, $d_i \geq 0, i = 1, 2$ 是惩罚参数. 首先利用 Lagrangian 方程及利用满足的 KKT 条件, 得到式(6)和式(7)表示的对偶问题

$$\max_{\alpha} \quad \alpha^T G (H^T H)^{-1} G^T \alpha \quad (8)$$

$$\text{s. t. } 0 \leq \alpha \leq d_1 e_-$$

$$\max_{\beta} \quad \beta^T P (Q^T Q)^{-1} P^T \beta \quad (9)$$

$$\text{s. t. } \leq \beta \leq d_2 e_+$$

其中,

$$G = [B e_-], H = [A e_-], P = [A e_+], Q = [B e_-],$$

$$e_+ = (1, \dots, 1)^T \in \mathcal{R}^p, e_- = (1, \dots, 1)^T \in \mathcal{R}^q,$$

$$A = (x_1, \dots, x_p)^T \in \mathcal{R}^{p \times n}, B = (x_{p+1}, \dots, x_{p+q})^T \in \mathcal{R}^{q \times n}$$

通过上述的计算可以求出式(6)和式(7)的解 (w_+, b_+) 和 (w_-, b_-) . 对于一个新的数据 $x \in \mathcal{R}^n$, 由式(10)判别其情感类别.

$$\text{Class} = \arg \min_{k=-, +} |(w_k \cdot x) + b_k| \quad (10)$$

式中, $|\cdot|$ 是点 x 到超平面 $(w_k \cdot x) + b_k = 0, k = -, +$ 的垂直距离.

TWSVM 提高了支持向量机的准确度, 同时提高了支持向量机的运行速度^[4], 但 TWSVM 运算过程中需要计算矩阵的逆, 其对于大规模数据处理依然很困难.

3.2 NPSVM

非平行支持向量机(TWSVM)是基于支持向量机和孪生支持向量机提出的, 主要目标也是找到两个超平面拟合各类别数据, 但由于其形式和标准的支持向量机十分相似, 因此可以用解标准的支持向量机的技术求解非平行支持向量机.

对于两类数据(如(5))找到两个非平行的超平

面 $(w_+ \cdot x) + b_+ = 0$ 和 $(w_- \cdot x) + b_- = 0$. 这两个超平面能被以下两个凸 QPPs 解决:

$$\begin{aligned} \min_{w_+, b_+, \eta_+^*, \xi_-} \quad & \frac{1}{2} (\|w_+\|^2) + C_1 \sum_{i=1}^p (\eta_i + \eta_i^*) + C_2 \sum_{j=p+1}^{p+q} \xi_j, \\ \text{s. t.} \quad & (w_+ \cdot x_i) + b_+ \leq \varepsilon + \eta_i, i = 1, \dots, p, \\ & -(w_+ \cdot x_i) - b_+ \leq \varepsilon + \eta_i^*, i = 1, \dots, p, \\ & (w_+ \cdot x_j) + b_+ \leq -1 + \xi_j, j = p+1, \dots, p+q, \\ & \eta_i, \eta_i^* \geq 0, i = 1, \dots, p, \\ & \xi_j \geq 0, j = p+1, \dots, p+q, \end{aligned} \tag{11}$$

$$\begin{aligned} \min_{w_+, b_+, \eta_+^*, \xi_-} \quad & \frac{1}{2} (\|w_-\|^2) + C_3 \sum_{i=p+1}^{p+q} (\eta_i + \eta_i^*) + C_4 \sum_{j=1}^p \xi_j, \\ \text{s. t.} \quad & (w_- \cdot x_i) + b_- \leq \varepsilon + \eta_i, i = p+1, \dots, p+q, \\ & -(w_- \cdot x_i) - b_- \leq \varepsilon + \eta_i^*, i = p+1, \dots, p+q, \\ & (w_- \cdot x_j) + b_- \geq -1 + \xi_j, j = 1, \dots, p, \\ & \eta_i, \eta_i^* \geq 0, i = p+1, \dots, p+q, \\ & \xi_j \geq 0, j = 1, \dots, p \end{aligned} \tag{12}$$

式中, $C_i \geq 0, i = 1, \dots, 4$ 是惩罚参数. 首先利用 Lagrangian 方程及利用满足的 KKT 条件, 得到式 (6) 和式 (7) 表示的对偶问题:

$$\begin{aligned} \min_{\alpha_+^*, \beta_-} \quad & \frac{1}{2} (\alpha_+^* - \alpha_+)^T A A^T (\alpha_+^* - \alpha_+) - \\ & (\alpha_+^* - \alpha_+)^T A B^T \beta_- + \frac{1}{2} \beta_-^T B B^T \beta_- + \end{aligned} \tag{13}$$

$$\begin{aligned} & \varepsilon e_+^T (\alpha_i^* + \alpha_i) - e_-^T \beta_-, \\ \text{s. t.} \quad & e_+^T (\alpha_+ - \alpha_+^*) + e_-^T \beta_- = 0 \\ & 0 \leq \alpha_+, \alpha_+^* \leq C_1 e_+, \quad 0 \leq \beta_- \leq C_2 e_-. \\ \min_{\alpha_-^*, \beta_+} \quad & \frac{1}{2} (\alpha_-^* - \alpha_-)^T B B^T (\alpha_-^* - \alpha_-) + \\ & (\alpha_-^* - \alpha_-)^T B A^T \beta_+ + \frac{1}{2} \beta_+^T A A^T \beta_+ + \end{aligned} \tag{14}$$

$$\begin{aligned} & \varepsilon e_+^T (\alpha^* + \alpha) - e_+^T \beta_+, \\ \text{s. t.} \quad & e_-^T (\alpha_- - \alpha_-^*) + e_+^T \beta_+ = 0 \\ & 0 \leq \alpha_-, \alpha_-^* \leq C_3 e_-, \quad 0 \leq \beta_+ \leq C_4 e_+. \end{aligned}$$

其中

$$\alpha_+^* = (\alpha_+^{*T}, \alpha_+^{*T}) = (\alpha_1, \dots, \alpha_p, \alpha_1^*, \dots, \alpha_p^*)^T,$$

$$\begin{aligned} \alpha_-^* &= (\alpha_-^{*T}, \alpha_-^{*T}) = (\alpha_p, \dots, \alpha_{p+q}, \alpha_p^*, \dots, \alpha_{p+q}^*)^T, \\ \beta_+ &= (\beta_1, \dots, \beta_p), \beta_- = (\beta_{p+1}, \dots, \beta_{p+q}). \end{aligned}$$

通过上述的计算可以求出式 (6) 和式 (7) 的解 (w_+, b_+) 和 (w_-, b_-) . 一个新的数据 $x \in \mathcal{R}^n$ 是通过式 (10) 来判断其为正向或负向.

$$\text{Class} = \arg \min_{k=-,+} | (w_k \cdot x) + b_k |.$$

非平行支持向量机在计算, 要比孪生支持向量机上要容易很多, 因此运行速度比后者更快, 也容易实现. 由于这两种支持向量机的主函数是成对出现的, 我们称之为双目标函数的支持向量机.

4 实验分析

本文主要采用的数据库是情感分析中常用的标准数据库. 第一个数据集是因特尔电影数据库 (Internet Movie Database (IMDB)). 第二数据集是 pang(2002) 的标准数据库. 这两个数据集的特点如表 1 所示. 在具体的实验中为了取得较好地精度执行效率, 对数据进行预处理, 选取文档频率大于 2 并且文档集频率大于 10 的情感词作为特征, 采用 TF-IDF 进行文本表示. 实验将每个数据集随机平分成训练集合测试集.

表 1 数据集的特点

数据集	样本大小	正向样本	负向样本	特征数
Pang2002	2 000	1 000	1 000	8 614
IMDB	4 000	2 000	2 000	5 706

实验将支持向量机 (libSVM 和 SMO) 与双目标函数支持向量机 (NPSVM 和 TWSVM) 在上述两个标准数据集上进行对比验证. 本实验发现在该数据库上用线性核函数比用径向基核函数效果好.

所以实验方法均采用线性核函数. NPSVM, TWSVM 中的参数 $C_i, i = 1, \dots, 4$ 是从 2^{-5} 到 2^5 范围内自动搜所最优参数的. 最优的参数 ε 是在 $[0, 0.5]$ 内以步长 0.05 取得的. 对于 libSVM 和 SMO 是采用了 WEKA 开源工具实现的. 实验结果如表 2、表 3 所示. 为了进一步说明算法的稳定性和准确度, 我们在 IMDB 数据集上对占不同百分比的训练集进行测试, 其实验结果如图 1 所示.

表 2 在标准数据集 Pang2002 实验结果 %

Methods	Accuracy
libSVM	79.8
SMO	81.3
TWSVM	82.5
NPSVM	82.8

表 3 在标准数据集 IMDB 实验结果 %

Methods	Accuracy
libSVM	88.3
SMO	86.1
TWSVM	90.6
NPSVM	89.5

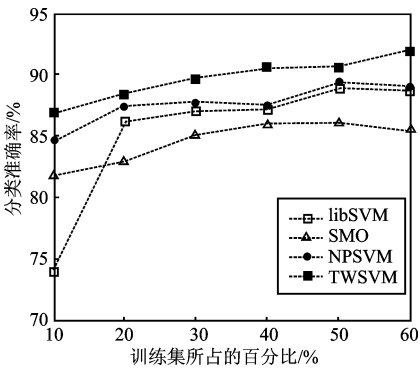


图 1 在标准数据集 IMDB 上平均结果

分析以上结果:从表 2 和表 3 可以看出双目标函数支持向量机在这两个数据库上运行分类精确度较 SVM 分类器要高 1%~4%. 从图 1 可明显地看出 NPSVM 和 RWSVM 的稳定性与准确度都要比 libSVM 和 SMO 分类器要好.

5 结束语

本文将双目标函数的支持向量机 NPSVM 和 TWSVM 应用到情感分析中,并取得较好地结果. 通过实验本文还发现双目标函数支持向量机尤其是 NPSVM 在运行速度上比支持向量机速度要快,并且具备精度高,适用的范围广的特点. 在情感分析中双目标函数支持向量机作为分类器较一般支持向量机拥有更强的泛化性能.

参考文献:

[1] 黄萱菁,张奇,吴苑斌. 文本情感倾向分析[J]. 中文信息学报, 2011, 25(6): 118-126.

[2] PangB, Lee L, Vaithyanath S. Thumbs up sentiment classification using machine learning technique [J]. Proc of the EMLP 2002,10(7):79-86.

[3] V. N. Vapnik, An overview of statistical learning theory[J]. IEEE Trans Neural Networks, 1999, 10(1): 988-999.

[4] Ayadeva R K, Khemchandani R, Chandra S. Twin support vector machines for pattern classification[J]. IEEE Trans Pattern Anal Mach Intell, 2011,29(5):905-910.

[5] Yingjie Tian, Xuchan Ju, Zhiquan Qi. Nonparallel support vector for pattern classification [J]. IEEE Trans Cybernetics, 2014, 44(7): 1067-1076.

[6] Tao Li, Ti Zhang, Vikas Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge[C]//proceedings of the Joint Conference of the 47th Annual Ueeting of the ACL and the 4th International Joint Conference on Natural Language Proceessing of the AFNLP. stroudsburg, PA, USA: IEEE, 2009: 244-252.

[7] Ahmed Abbasi, Hsinchun Chen, Arab Salem. Sentiment analysis in multiple languages; feature selection for opinion classification in Web forums[J]. Journal ACM Transactions on Infarmation Systems Tois Homepage archive, 2008, 26(3): 1-12.

作者简介:

刘春雨 女, (1988-), 硕士. 研究方向为应用数学.
E-mail: 1278678839@qq. com

郭满才 男, (1963-), 教授. 研究方向为应用统计.

朱 倩 女, (1993-), 硕士. 研究方向为自然语言.

周晓飞 女, (1978-), 研究员. 研究方向为机器学习.