

一种基于支持向量机的半监督分类方法

徐庆伶, 汪西莉

(陕西师范大学 计算机科学学院, 陕西 西安 710062)

摘 要: 如何有效利用海量的数据是当前机器学习面临的一个重要任务, 传统的支持向量机是一种有监督的学习方法, 需要大量有标记的样本进行训练, 然而有标记样本的数量是十分有限的并且非常不易获取。结合 Co-training 算法与 Tri-training 算法的思想, 给出了一种半监督 SVM 分类方法。该方法采用两个不同参数的 SVM 分类器对无标记样本进行标记, 选取置信度高的样本加入到已标记样本集中。理论分析和计算机仿真结果都表明, 文中算法能有效利用大量的无标记样本, 并且无标记样本的加入能有效提高分类的正确率。

关键词: 半监督学习; 支持向量机; 遗传算法

中图分类号: TP181

文献标识码: A

文章编号: 1673-629X(2010)10-0115-03

A Novel Semi-Supervised Classification Method Based on SVM

XU Qing-ling, WANG Xi-li

(School of Computer Science, Shaanxi Normal University, Xi'an 710062, China)

Abstract: One of the important assignment in machine learning is how to use large-scale data effectively, the traditional SVM is a kind of supervised learning approach, it needs a number of labeled samples for training, but the labeled samples are limited and very difficult to obtain. A semi-supervised SVM for classification is proposed by binding the thoughts of Co-training and Tri-training together. This method uses two SVM classifiers with different parameters to label the unlabeled samples, then chooses the samples with high confidence level to extend the labeled sample-set. Both theoretical analysis and simulation results indicate that this method can use a lot of unlabeled samples effectively, and the addition of unlabeled samples can improve classification accuracy available.

Key words: semi-supervised learning; support vector machine(SVM); genetic algorithm(GA)

0 引言

机器学习问题可以分为监督学习和无监督学习两种, 传统的非监督学习虽然算法成熟, 操作简单, 但其学习效果往往很难得到保证; 监督学习则需要大量的已标记的学习样本, 在很多实际应用中, 获取大量的无标记样本非常容易, 而获取有标记的样本通常需要付出较大的代价, 新型半监督学习结合了监督学习与非监督学习的优点, 适合于已标记样本较少, 同时具有大量未标记样本的分类问题^[1,2]。按照 Chapelle、Seeger、Mitchell 等人对半监督学习的分析, 半监督学习问题主要有三种主要的技术: 基于生成式的模型、基于图正则化框架的模型以及基于协同训练的模型^[3]。

由 V. Vapnik 创立的支持向量机(SVM)^[4]建立在

统计学习理论和结构风险最小化原则之上, 具有非常坚实的理论基础。而且该模型需要设定的参数相对较少, 非常适合于小样本机器学习, 目前已广泛应用于模式识别的分类器设计中^[5,6]。

文中将以协同训练(Co-training)模型为基础, 构建基于遗传算法(GA)优选参数的支持向量机(SVM)的半监督分类方法。本方法既融合了半监督方法能有效利用大量未标记样本的优点又具有 SVM 适用于进行小样本学习的特点, 并且通过理论及实验分析, 采用了半监督策略的支持向量机分类方法能够有效提高分类精度。

1 基于支持向量机的分类

支持向量机是建立在统计学习理论的 VC 维理论和结构风险最小化(SRM)原理基础上的新型学习机器^[7,8], 它是 SRM 原则的具体体现, 更是整个统计学习理论体系中最直观、最实用的部分。它根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷, 以期获得最好的推广能力^[4,9]。

收稿日期: 2010-01-27; 修回日期: 2010-05-08

基金项目: 国家自然科学基金(40671133)

作者简介: 徐庆伶(1984-), 女, 硕士研究生, 研究方向为智能信息处理、模式识别、图像处理; 汪西莉, 教授, 硕士生导师, 研究方向为智能信息处理、模式识别、图像处理等。

SVM 分类是从线性可分情况下的最优超平面发展起来的,它通过最大化分类间隔控制学习机器的容量来实现 SRM 原则^[10]。对于线性可分的两类问题,可直接构造最优分类面,使得样本集中的所有向量满足:

①能被某一超平面正确划分,这是为了保证经验风险最小化;

②距该超平面最近的异类向量与超平面之间的距离最大,即分类间隔最大,这是使 VC 置信度最小,从而使期望风险最小。这实际上是一个二次规划的问题,得到的最优决策函数如公式(1)所示:

$$F(x) = \text{sgn}\left\{\sum_{i=1}^L y_i a_i K(x_i, x) + b\right\} \quad (1)$$

其中, $K(\cdot, \cdot)$ 是核函数, $\text{sgn}(\cdot)$ 是符号函数, L 为训练样本的数目。系数 a_i 为公式(2)所示的二次优化问题的解:

$$\min_a \frac{1}{2} a^T Q a - e^T a \quad (2)$$

$$y^T a = 0$$

$$0 \leq a_i < \infty, i = 1, \dots, L \text{ (广义最优分类面 } 0 \leq a_i \leq C)$$

这里, Q 是一个 $L \times L$ 的半正定矩阵, $Q_{ij} = y_i y_j K(x_i, x_j)$, e 是所有元素都为 1 的列向量, C 为广义最优分类面而引入的误差惩罚因子。对于线性问题, $K(\cdot, \cdot)$ 就是两向量的点积运算。

而对于非线性问题, SVM 实现是通过某种事先选择的非线性映射(核函数)将输入向量 x 映射到某个高维特征空间 H , 在 H 中构造最优分类超平面。引入非线性映射 Φ 后, 原低维空间上的线性不可分问题就转化为一个高维特征空间上的线性可分或几乎线性可分问题, 将分类问题转换到特征空间上求解。引入一个核函数 $K(x, x_i) = \langle \Phi(x), \Phi(x_i) \rangle$ 对 Φ 进行整体处理, 这样可以避免对 Φ 的直接运算, 使得所有的运算仍在原空间进行。

目前应用较多的核函数主要有三类, 由于存在多种核函数, 设计 SVM 的主要内容之一便是选择核函数和核参数。Vapnik 等人的研究发现, 核函数的类型对 SVM 性能的影响不大, 而核参数和惩罚因子 C 是影响 SVM 性能的关键因素。因此, 核参数和惩罚因子 C 的选择对 SVM 的性能至关重要。目前, 径向基函数 SVM 的使用最为广泛, 文中对它的核参数 σ^2 和惩罚因子 C 进行优选。

误差惩罚因子 C 用于在确定的特征空间中调节学习器的置信范围和经验风险的比例, C 越大, 对训练样本数据的拟合程度越高, 但泛化能力将降低。核参数

σ^2 主要影响样本数据在高维特征空间中分布的复杂程度, 只有选择了合适的 σ^2 , 才能将数据映射到合适的特征空间中去。在 SVM 的具体应用中, 惩罚因子 C 和核参数 σ^2 对分类结果的影响均很大, 只有选择合适的模型参数, SVM 的优越性才能更好地发挥出来^[11,12]。因此, 文中采用遗传算法(GA)对 SVM 模型的参数进行优选。

2 基于支持向量机的半监督分类方法

最初的协同训练算法是 Blum 和 Mitchell 在 1998 年提出的, 标准协同训练(Co-training)算法假设数据有两个不同的充分冗余视图, 即属性集能够分为两个不相交的子集, 且每个属性子集都能独立训练出分类器^[3]。在这两个属性子集的基础上分别利用少量的已标记数据训练两个分类器, 再用训练得到的两个分类器分别对未标记样本进行预测, 并从中挑选出置信度较高的样本加入到训练集中, 以帮助对方重新训练分类器, 以改善性能。

然而, 在实际问题中充分冗余视图这一要求很难被满足, 所以 Goldman 和 Zhou 提出了一种 Co-training 的改进算法, 此算法不再需要充分冗余视图, 取而代之的是利用两个不同类型的分类器完成学习, 但要求在每轮迭代中采用 10 重交叉验证, 以确定未标记样本的标记置信度, 因此该方法十分耗时。针对 Co-training 及其改进算法存在的问题, 2005 年, Zhou 和 Li 提出了一种新的 Co-training 模式半监督分类算法 Tri-training^[13]。Tri-training 算法使用 3 个分类器进行训练, 其对属性集和 3 个分类器所采用的监督学习算法都没有约束。该算法的训练过程为: 假设初始标记样本集为 L , 由 L 训练得到 3 个不同的初始分类器 h_1 , h_2 和 h_3 , x 是无标记样本集 U 中的任意一个样本, 如果 h_2 和 h_3 对 x 的分类结果 $h_2(x)$ 和 $h_3(x)$ 一致, 那么就将 x 及其标记 $h_2(x)$ 加入到 h_1 的训练集中, 如此形成 h_1 的新训练集。类似地, h_2 和 h_3 的训练集也分别扩充为 S_2^1 和 S_3^1 , 然后 3 个分类器重新训练, 如此重复迭代直至 h_1 , h_2 和 h_3 都没有变化, 训练过程结束。但 Tri-training 算法中存在大量反复训练过程, 所以其会耗费一定时间。

文中结合 Co-training 算法与 Tri-training 算法的思想, 提出一种简化的算法, 该算法在同一个训练集上训练得到两个不同的分类器, 然后用这两个分类器对无标记样本进行预测, 选取置信度高的无标记样本及其预测标记加入到训练集中, 更新训练集, 重新训练分类器, 如此循环上述步骤, 直至满足停止条件。这样既避免了 Co-training 算法要求属性集拥有两个充分冗

余视图的假设,又大大简化了 Tri-training 训练过程中需要三个分类器进行重复迭代的过程,算法具体描述如下:

假设样本集为: $S = L \cup U$, 其中 $L = \{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i)\}$ 为已标记样本, 并将 L 分为 L_1 、 L_2 两部分 ($L = L_1 \cup L_2$, L_1 为训练样本集, L_2 为测试样本集), $U = \{(x_{i+1}, x_{i+2}, \dots, x_n)\}$ 为无标记样本。

算法流程如下:

- ① 输入已标记样本集 L_1 , 未标记样本集 U ;
- ② 用 L_1 分别训练分类器 h_1, h_2 ;
- ③ 用 h_1 对 U 进行标记, 得到结果为 u_1 ;
- ④ 用 h_2 对 U 进行标记, 得到结果为 u_2 ;
- ⑤ 比较 u_1 和 u_2 , 将 u_1 和 u_2 标记结果一致的样本记为 u' ;

⑥ 若未达到循环终止条件, 则将 u' 加入到 L 中, 返回 ②; 若达到了循环终止条件, 则退出循环, 用最终得到的训练样本集 L 训练分类器, 选取 h_1, h_2 中分类结果好的一个对测试样本进行测试。

算法中, h_1 和 h_2 需为不同类型的分类器, 否则协同训练算法则退化成了自训练算法, 文中 h_1 和 h_2 虽均为 SVM 分类器, 它们的差异来源于其所选取的参数, 其中 h_1 所用参数为默认值, h_2 所选参数则为经过遗传算法优选出来的参数。在 L 上训练得到的两个分类器对同一未标记样本集 U 进行预测, 直观上认为两次标记结果一致的样本具有较高的置信度, 因此可将其加入到 L 中用以扩大训练样本集并重新训练分类器。

3 实验结果与分析

所用实验数据均选自 UCI, 所有数据的类别均为已知, 具体操作中, 已标记样本从数据集中直接选取, 无标记样本则需手工去除掉类标记。所用实验数据集如表 1 所示。

表 1 实验所用数据

数据集	样本个数	属性个数	类别数
Yeast	1481(1394)	8	10(4)
Iris	150	4	3
Ionosphere	351	34	2

其中 Yeast 数据集的类别按样本个数排列为: CYT、NUC、MIT、ME3、ME2、ME1、EXC、VAC、POX、ERL。为了便于实验, 将 EXC、VAC、POX、ERL 四类个数较少的样本全部去掉, 并将 ME3、ME2、ME1 三类样本标记为一类, 则实验数据总共为四类。

由于实验所用训练样本及测试样本均为从数据集

中随机选取出来的, 为了去除实验结果的随机性, 最终记录的实验结果为三次实验所得结果的平均值。其中 SVM 方法采用 OUS-SVM3.0 工具箱, 核函数选取为 RBF 核函数, 参数均为默认值, 其中惩罚系数 C 设为 1, RBF 核参数 $\sigma^2 = 1$; GA-SVM 方法中 SVM 所采用的参数是经过遗传算法优选所得; Semi-svm 方法则为基于半监督策略的 SVM 与 GA-SVM 协同进行分类的方法。其中 Yeast 数据集中训练样本由每类中抽取 20 个样本组成, 测试样本由每类中抽取 10 个样本组成, 剩余的 1274 个样本组成无标记样本; Iris 数据集中训练样本及测试样本分别为从每类中抽取出来 10 个样本组成, 剩余 90 个样本组成无标记样本; Ionosphere 数据集中训练样本及测试样本分别为从每类中抽取 10 个样本组成, 剩余 311 个样本组成无标记样本。

实验结果如表 2 所示。

表 2 分类结果比较

分类方法 正确率 实验数据	SVM	GA-SVM	Semi-svm
Yeast	0.70	0.80	0.83
Iris	0.90	0.97	1.00
Ionosphere	0.80	0.90	0.95

从表 2 可以看出, 在分类正确率方面, 应用了遗传算法优选参数的支持向量机的分类结果明显优于选用默认参数的支持向量机的分类结果, 而基于半监督策略的 SVM 与 GA-SVM 协同进行训练的方法所得到的结果比较前两种方法有所提高, 这说明采用了半监督策略的方法能够通过无标记样本的加入使分类精度提高。

4 结束语

文中结合 Co-training 算法与 Tri-training 算法的思想, 给出了一种半监督 SVM 分类方法, 采用的方法能有效利用大量的无标记样本, 并且从实验结果可以看出, 无标记样本的加入有效提高了分类的正确率。但是由集成策略可知, 分类器间的差异越大, 最后的集成效果也越好, 因此下一步可以考虑选取不同类型的分类器, 以期获得更好的分类效果。

参考文献:

- [1] Zhu X J. Semi-supervised learning literature survey[R]. U. S. A: University of Wisconsin-Madison, 2005.
- [2] 易 星. 半监督学习若干问题的研究[D]. 北京: 清华大学,

(下转第 121 页)

仿真结果显示,优化后的 Flooding 具有较长的网络存活时间和较高的网络利用效率。同时,由于其策略简单、有效,可以在现有的路由协议上快速扩充,实用性较强。将其扩充到能量受限的 QoS 单播和多播路由协议^[17]里面,也是当前工作的一个方向。

参考文献:

- [1] 徐雷鸣, 庞博, 赵耀. NS 与网络模拟 [M]. 北京: 人民邮电出版社, 2003.
- [2] Macker J, Chakeres I. Mobile Ad-hoc Networks (manet) [EB/OL]. 2008-10-13. <http://www.ietf.org/html.charters/manet-charter.html>.
- [3] Grossglauser M, Tse D N C. Mobility increases the capacity of ad hoc wireless networks[J]. IEEE/ACM Transactions on Networking (TON), 2002, 10(4): 477-486.
- [4] Tseng Y C, Ni S Y, Chen Y S, et al. The Broadcast Storm Problem in a Mobile Ad Hoc Network[J]. ACM Wireless Networks, 2002, 8(2): 153-167.
- [5] Navarra A. 3-Dimensional minimum energy broadcasting problem[J]. Ad Hoc Networks, 2008, 6(5): 734-743.
- [6] 袁培燕, 崔金玲. 一种能量负载均衡的自组织网络多播路由协议[J]. 河南师范大学学报: 自然科学版, 2008, 36(6): 37-39.
- [7] Tseng Y C, Ni S Y, Shih E Y. Adaptive approaches to relieving broadcast storms in a wireless multihop mobile ad hoc networks[J]. IEEE Transactions on Computers, 2003, 52(5): 545-557.
- [8] Parkinson B, Spilker J. Global Positioning System: Theory and Application[M]. US: American Institute of Astronautics and Aeronautics, 1996.
- [9] Ko Y B, Vaidya N H. Location-aided routing (LAR) in mobile ad hoc networks[C]//in: ACM MOBICOM, 1998. [s.l.]: [s.n.], 1998.
- [10] Niculescu D, Nath B. Position and orientation in ad hoc networks[J]. Ad Hoc Networks, 2004, 2(2): 133-151.
- [11] Bose P, Morin P, Stojmenovic I, et al. Routing with guaranteed delivery in ad hoc wireless networks[J]. Wireless Networks, 2001, 7(6): 609-616.
- [12] Kranakis E, Singh H, Urrutia J. Compass routing on geometric networks[C]//in: Proceedings of 11th Canadian Conference on Computational Geometry. [s.l.]: [s.n.], 1999: 51-54.
- [13] Moaveninejad K, Song Wen-Zhan, Li Xiang-Yang. Robust position-based routing for wireless ad hoc networks[J]. Ad Hoc Networks, 2005, 3(1): 546-559.
- [14] Intanagonwiwat C, Govindan R, Estrin D. Directed diffusion: A scalable and robust communication paradigm for sensor networks[C]//In: Pickholtz R. Proc. of the 6th Annual Int'l Conf. on Mobile Computing and Networking. New York: ACM Press, 2000: 56-67.
- [15] Heidemann J, Silva F, Estrin D. Matching data dissemination algorithms to application requirements[J]. In: Akyildiz I F. Proc. of the 1st Int'l Conf. on Embedded Networked Sensor Systems. New York: ACM Press, 2003: 218-229.
- [16] Camp T, Boleng J, Davies V. A survey of mobility models for ad hoc networks research[J]. Wireless Communication and Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications, 2002, 2(5): 483-502.
- [17] Li Layuan, Li Chunlin. A QoS-guaranteed multicast routing protocol[J]. Computer Communications, 2004, 27(1): 59-69.

(上接第 117 页)

- [1] 周志华. 机器学习及其应用 [M]. 北京: 清华大学出版社, 2007: 259-275.
- [2] Vapnik V. The Nature of Statistical Learning [M]. New York: Springer, 1995.
- [3] Ge M, Du R, Zhang C C, et al. Fault diagnosis using support vector machine with an application in sheet metal stamping operations[J]. Mechanical Systems and Signal Processing, 2004, 18: 143-159.
- [4] Guo G D, Li S Z. Content-based Audio Classification and Retrieval by Support Vector Machines[J]. IEEE Trans. on Neural Network, 2003, 14(1): 209-215.
- [5] 边肇祺, 张学工. 模式识别 [M]. 北京: 清华大学出版社, 2000.
- [6] Gunn S R. Support Vector Machines for Classification and Regression[R]. Britain: University of Southampton, 1997.
- [7] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods [M]. Beijing: Publishing House of Electronics Industry, 2004.
- [8] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.
- [9] 杨旭, 纪玉波, 田雪. 基于遗传算法的 SVM 参数选取[J]. 辽宁石油化工大学学报, 2004, 24(1): 54-58.
- [10] 周兆永, 汪西莉, 曹艳龙. 基于 GA 优选参数的 SVM 水质评价方法研究[J]. 计算机工程与应用, 2008, 44(4): 190-193.
- [11] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.