

基于数据相似度的自适应半监督随机森林算法

胡志鹏, 彭亦功

(华东理工大学 信息科学与工程学院, 上海 200237)

摘 要: 提出一种基于数据相似度的自适应半监督随机森林算法. 利用随机森林对带标签和无标记数据进行路径编码、相似度分析和无标签数据的伪标记选择; 再选择满足条件的数据迭代训练随机森林, 改善其分类性能. 实验结果表明: 提出的算法可以有效地利用无标记数据信息, 提高分类精度.

关键词: 随机森林; 半监督学习算法; 数据相似度; 路径集合稀疏编码; 自适应

中图分类号: TP312

文献标识码: A

文章编号: 1000-7180(2018)07-0117-05

Adaptive Semi-Supervised Random Forest Algorithm Based on Data Similarity

HU Zhi-peng, Peng Yi-gong

(College of Information Science and Engineering, East China University of Science and Technology
Shanghai 20023, China)

Abstract: An adaptive semi-supervised stochastic forest algorithm based on data similarity is proposed. The coding and mapping of labeled and unlabeled data are carried out by using random forest. The pseudo-marking of unlabeled data is selected. Then the randomized forest is trained to improve the classification performance. The experimental results show that the proposed algorithm can effectively use the unmarked data to improve the classification accuracy.

Key words: random forests; semi-supervised algorithm; data similarity; path set sparse coding; adaptive

1 引言

随机森林(RF)是一种组合学习算法^[1], 被广泛应用于模式识别, 但在训练数据集中、数据标签不足对随机森林的构建是一个严峻的挑战; 因为使用少量带标签数据构建随机森林将会导致算法性能的明显下降^[2]. 并且, 标记数据是费时和昂贵的. 克服上述问题的主要解决方案是引入未标记数据来指导分类器学习, 也就是半监督学习算法(SSL). 基于边缘最大的 SSL 方法的有效性在很大程度上取决于数据的分布; 通常情况下, 训练数据难以满足要求^[2]; 基于节点划分的半监督随机森林算法^[2]: 利用超倾斜平面对随机森林中决策树叶节点进行分割, 依据概率分布估计未标记数据的标签, 再使用袋外样

本^[1]选择标记数据训练随机森林; 一种用于计算机辅助诊断的新型协同训练随机森林^[3]: 该算法在每次迭代中放弃置信度偏低的伪标记数据, 提高了伪标记精度, 改善了分类器的性能.

基于上述研究, 本文提出一种基于数据相似度的自适应半监督随机森林算法(ad-RF).

2 基于数据相似度的自适应半监督随机森林

2.1 路径集合稀疏编码

由于实例在决策树中会生成一条清晰的反映数据样本相似度的遍历路径, 假设历经的节点编码 1, 其余节点编码 0, 定义产生的二进制序列为路径集合稀疏编码, 编码长度是随机森林总节点数. 在此过程中, 随机树作用是空间划分而不是分类器; 通过将初始样本特

征投影到高维空间,实现了样本分类边界的清晰化;并且,编码过程是数据信息的表达过程,包含了样本的局部特征价值信息和判别信息.假设某个实例在森林中的遍历路径如图 1 虚线所示,则生成的编码是 1101000 1101000 1101000.

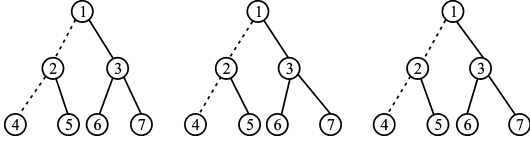


图 1 实例在决策树节点历经示意图

2.2 基于 Jaccard 准则的相似度量

假设随机森林的规模是 M ,实例 x 和 z 的每条路径集合稀疏编码中相同节点个数是 a_i ,不同节点个数是 b_i ;相同叶节点个数是 c .根据 jaccard 准则度量两个实例的相似度 $S(x, z)$ 如式(1)所示,文献[4]度量准则如式(2)所示:

$$S(x, z) = 1 - \frac{(\sum_{i=1}^M \frac{b_i}{a_i + b_i})}{M} \quad (1)$$

$$S(x, z) = \frac{c}{M} \quad (2)$$

公式(1)的度量准则与公式(2)的讨论如下:

假设 $c = \lambda M$ 即有不等式(3)成立:

$$1 - \frac{\sum_{i=1}^{(1-\lambda) \cdot M} \frac{b_i}{a_i + b_i}}{M} \geq \lambda \quad (3)$$

综上所述:对比文献[3]度量准则发现:Jaccard 准则能够提供清晰的分类边界,有利于提高伪标记精度和进一步改善分类器的性能,仿真结果在 4.3 节.

3 伪标记和自适应噪声补偿

通过第 1 节的 Jaccard 准则计算实例相似度,选择相似度平均值最大的类标签,标记无标签数据.若 C 类的伪标记数据个数是 U_c ,带标签数据个数是 L_c ,未标记样本 x 与 C 类带标签数据的平均相似度 S 如公式(4):

$$S = \frac{\sum_{i=1}^{L_c} RF(x, x_i)}{L_c} \quad (4)$$

未标记数据通过公式(4)选择伪标记.假设同类伪标记数据与同类带标签数据平均相似度是 AS ,置信参数是 α ;改变 α 可以控制伪标记精度和数量;

如果伪标记实例的相似度量低于阈值,即认为伪标记实例被错误标记;尽管 Jaccard 准则能够提高伪标记精度,但伪标记集合中仍然会存在噪声,影响分类器性能.受文献[2]启发,若加入的伪标记数据能够满足一定条件,即可补偿错误标记数据对分类器性能的影响,所需满足条件如公式(5)所示:

$$m \geq \frac{2}{\beta^2 (1 - 2\theta)^2} \ln\left(\frac{2H}{\sigma}\right) \quad (5)$$

式中, m 是第 t 次迭代中的数据个数; β 是分类器错误率上限; θ 是分类器噪声率上限; H 是当次迭代中伪标记数据数量; σ 是分类置信度.为了使(6)式等号成立,等式右边乘以一个等价因子 μ ,且令 $d = 2\mu \ln\left(\frac{2H}{\sigma}\right)$;得到公式(7):

$$m = \frac{d}{\beta^2 (1 - 2\theta)^2} \quad (6)$$

$$u = \frac{d}{\beta^2} = m (1 - 2\theta)^2 \quad (7)$$

假设第 $t-1$ 次迭代过程中伪标记数据个数是 L^{t-1} ,即带标签的数据总数是 $L \cup L^{t-1}$,伪标记错误率是 e^{t-1} ;若第 t 次迭代过程中伪标记数据个数是 L^t ,带标签的数据总数是 $L \cup L^t$;伪标记错误率是 e^t .以第 t 次迭代的伪标记错误率计算为例:如公式(8), $count^{right}$ 表示正确伪标记数据个数.值得注意的是: $t-1$ 时刻被标记的数据不会并入原始带标签数据中,且在第 t 次迭代中仍然是无标记数据,然后被伪标记.假设 η 是原始标签集合的噪声率,在第 t 次迭代和 $t-1$ 次迭代中的噪声率如公式(9)和公式(10):

$$e^t = \frac{count^{right}}{L^t} \quad (8)$$

$$\theta^t = \frac{\eta |L| + e^t |L^t|}{|L \cup L^t|} \quad (9)$$

$$\theta^{t-1} = \frac{\eta |L| + e^{t-1} |L^{t-1}|}{|L \cup L^{t-1}|} \quad (10)$$

将公式(9)和(10)分别代入式(7)中,得公式(11)和(12):

$$u^t = |L \cup L^{t-1}| \left(1 - 2 \frac{\eta |L| + e^t |L^t|}{|L \cup L^t|}\right)^2 \quad (11)$$

$$u^{t-1} = |L \cup L^{t-1}| \left(1 - 2 \frac{\eta |L| + e^{t-1} |L^{t-1}|}{|L \cup L^{t-1}|}\right)^2 \quad (12)$$

由 $u \propto \frac{1}{\beta^t}$ 和公式(11) 和(12) 可知,若 $u^t > u^{t-1}$,即 $\beta^t < \beta^{t-1}$;则说明加入伪标记数据降低了分类器的错误率.通过对上述不等式讨论,得出所需满足条件如公式(13) 所示.该条件决定是否使用当次伪标记的数据

$$0 < \frac{e^t}{e^{t-1}} < \frac{|L^{t-1}|}{|L^t|} < 1$$

(13)

由于估计无标记数据的标签准确性是困难的,因此在标签数据和未标记数据是同构的假设前提下,利用原始标签中数据评估伪标记精度,通过调整置信参数,实现噪声自适应补偿.算法流程如下:

输入:带标签数据集 L
无标记数据集 U
学习算法:随机森林
参数 α 和迭代步长 c

$RF \leftarrow training(L), Spare \ coding \leftarrow transform(|L \cup U|)$

迭代更新,直至 RF 分类器性能不发生改变

① 计算已编码的未标记数据与带标签数据的相似度

② 调整参数 $\alpha \leftarrow \alpha + c$

③ 提取相似度大于 $\alpha(AS)$ 的伪标记数据

④ 评估提取的伪标记数据错误率是否满足 $e^t < e^{t-1}$,若成立,执行步骤 ⑤;若不成立,跳出本次循环

⑤ 计算数据数量是否满足公式(17) 条件,若成立,执行步骤 ⑦;若不成立,执行步骤 ⑥

⑥ 采用有放回随机抽样,执行步骤 ⑤

⑦ 使用原始带标签数据 L 和步骤 ④ 得到的伪标记数据,共同训练 RF

⑧ 使用袋外样本对 RF 进行性能评估

输出:
for $\in U$

$y_i = majority \ vote \ (F_R)_j^{N_{tree}} = 1$

4 实验

4.1 数据集

手写体数据集来源是著名的机器学习竞赛网站 KAGGLE,Iris^[6] 和 Wine 数据集来源是 UCI 机器学习数据库^[5].由于原始手写体数据集规模较大,所以进行了筛选.三个数据集的属性、类别、规模如表 1 所示.

分别对数据集进行无放回随机抽样,抽样比分别是 5%,10%,15%和 20%,所得抽样数据作为训练集;剩余数据作为测试集.

表 1 数据集

名称	来源	属性	类别	数据集规模
手写体数据集	KAGGLE	784	10	900
Iris 数据集	UCI	5	3	150
Wine 数据集	UCI	13	3	178

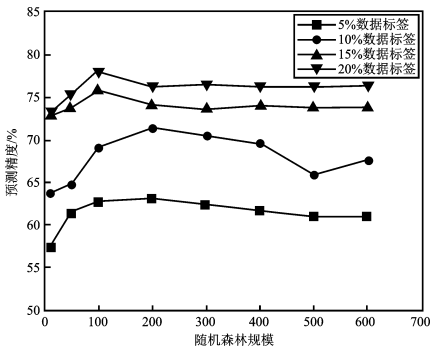
4.2 参数选择和算法复杂度分析

在本文的算法中,需要给定的参数是随机森林的规模、超参数 α 和迭代步长 c .迄今为止的研究中,无法准确给定最优森林规模,本文算法的森林规模确定是通过仿真实验得到,选择 200;图 2(a),(b) 和(c)显示了森林规模对算法精度的影响.值得注意的是:对比算法的参数通过上述方法确定;另一方面,初始超参数和迭代步长的确定是根据数据相似度的分布,手写体数据集中超参数是 1.00,迭代步长是 0.01;Iris 数据集中超参数是 0.85,迭代步长是 0.05;Wine 数据集中超参数是 0.95,迭代步长是 0.02;算法错误率均值的迭代曲线如图 3 所示;基于文献[3]的相似度量和本文相似度量对算法精度的影响曲线如图 4(a)和(b)所示.通过分析可知,

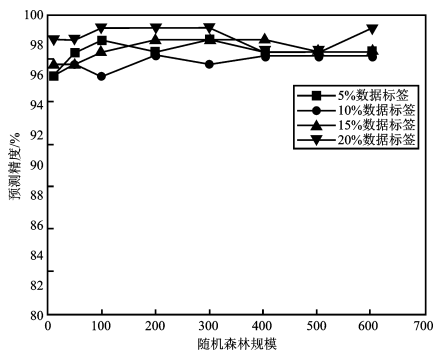
算法时间复杂度是: $O(TmaxNK)$,其中 $Tmax$ 是森林规模, N 是样本数, K 是叶子节点个数.值得注意的是:稀疏编码有着高效的存储和计算速度,并且,随机森林算法适合并行部署.

4.3 仿真和结果对比

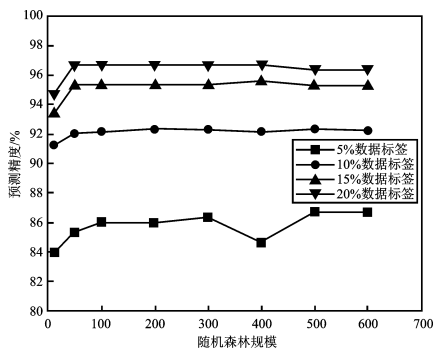
为了评估本文算法的有效性和可行性及其相对于其他算法的较高的分类精度,本文算法分别和随机森林算法、极端森林^[8]算法、自学习随机森林算法和协同^[9]随机森林算法,在三个不同数据集上进行了测试.算法实验环境是 Python 3.5,所用机器是笔记本(ACER V5-471 1.8GHz 4GB RAM),操作系统是 Windows 7 旗舰版.



(a) 手写体数据集



(b) Iris 数据集



(c) Wine 数据集

图 2 森林规模对算法 影响曲线

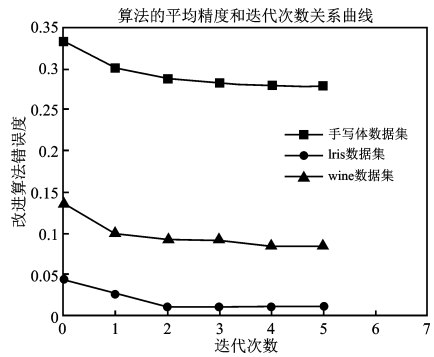
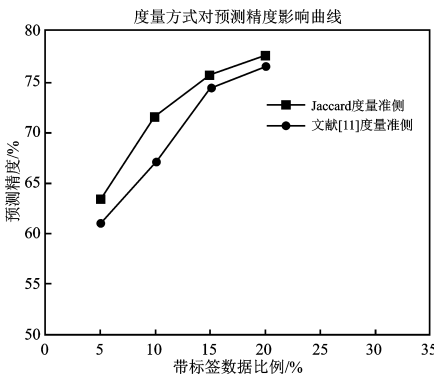
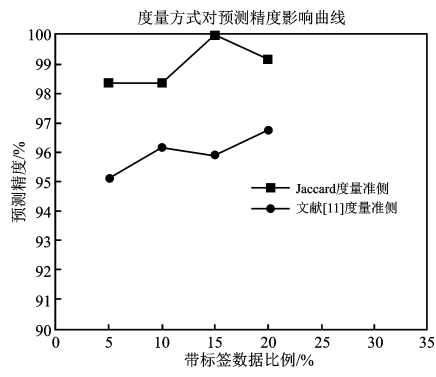


图 3 分类错误率迭代曲线



(a) 手写体数据集



(b) Iris 数据集

图 4 度量方式对预测精度影响曲线

4.4 结果分析

(1)对比初始随机森林算法,可知:本文算法的分类正确率在三个数据集上,平均改善了 5.52%,3.41%和 6.25%;自学习随机森林分类正确率平均改善了 2.46%,0.78%和 1.34%;协同随机森分类正确率平均改善了一0.48%,0.29%和 0.18%;可知:无论是在标准的 UCI 数据集上还是在分类难度更大的 KAGGLE 竞赛数据集上,本文算法都很好地适应小规模带标签样本数据环境,提升了分类正确率;

(2)对比初始随机森林算法,自学习随机森林算法在三个数据集上的分类精度都有改善,分别是 2.46%,0.78%和 1.07%;但是提升幅度受限.出现这种现象的原因是:在小规模带标签数据环境下,伪标记过程产生了过多错误标记,并且在迭代过程累积,使得算法的性能受限;协同随机森林算法在手写体数据集上的预测精度说明了这一点:大量错误标记的累积,恶化了分类器的性能,甚至导致算法预测精度的降低;在 Iris 和 Wine 数据集上,协同随机森林算法的预测精度是有所改善的,这种现象的原因是:标准数据集能够提供清晰的分类边界和随机森林算法出色的分类性能,故极大程度上弱化了噪声影响;

(3)随着带标签数据比例的增加,算法的预测精度提高;说明了数据信息在随机森林模型构建中的重要作用;

(4)由图 3 可知,手写字数据集和 Wine 四次迭代后收敛,Iris 迭代两次收敛;充分保证了改进算法的实时性;并且,随着随机森林规模的增加,分类精度得到提升,到达一定规模后稳定.因此在模型构建

过程中,需要选择适合的随机森林规模;并且,由图 4 可知,说明 Jaccard 的有效性和伪标记精度对算法性能的重要性.

5 结束语

本文提出的基于数据相似度的自适应半监督随机森林算法在标签数据不足情况下,通过稀疏编码的伪标记选择和自适应噪声补偿,降低了原始空间噪声影响和标记过程中错误标记影响,从而在一定程度上提高了随机森林算法的分类精度.并且,自适应迭代方法能够运用在数据标签标定环节,提升标定效率.未来会将研究该改进算法在入侵检测系统中运用.

参考文献:

- [1] Breiman L. Random forests[J]. Machine learning, 2001,5(1):5-32.
- [2] Liu X, Song M, Tao D, et al. Semi-supervised node splitting for random forest construction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013:492-499.
- [3] Deng C, Guo M Z. A new co-training-style random forest for computer aided diagnosis[J]. Journal of Intelligent Information Systems, 2011, 36(3):253-281.
- [4] Leistner C, Saffari A, Santner J, et al. Semi-Super-

vised Random Forests [C]// IEEE, International Conference on Computer Vision. IEEE, 2010: 506-513.

- [5] 陶剑文, Fu-Lai CHUNG, 王士同, 等. 稀疏标签传播:一种鲁棒的领域适应学习方法[J]. 软件学报, 2015, 26(5):977-1000.
- [6] Schölkopf B, Platt J, Hofmann T. Fast Discriminative Visual Codebooks using Randomized Clustering Forests[C]// International Conference on Neural Information Processing Systems. MIT Press, 2006: 985-992.
- [7] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data[C]//ICML. 2000: 327-334.
- [8] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1):3-42.
- [9] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]// Eleventh Conference on Computational Learning Theory. ACM, 1998: 92-100.

作者简介:

胡志鹏 男,(1993-),硕士研究生.研究方向为机器学习、模式识别、大数据.

彭亦功(通讯作者) 女,(1963-),博士,副教授.研究方向为物联网,人工智能等.

E-mail:cise_ecust@126.com.