

基于直觉模糊集的 Tri-Training 改进算法

彭雅琴¹, 宫宁生²

(1 三江学院 计算机科学与工程学院, 江苏 南京 210012;

2 南京工业大学 电子与信息工程学院, 江苏 南京 210009)

摘 要: Tri-Training 算法是半监督算法中的一种, 其初始分类器性能受有标记样本影响较大, 当样本数目不足时, 分类器性能相对较弱, 会直接影响后续迭代. 为此提出 IFS-Tri-Training (Tri-Training based on intuitionistic fuzzy sets) 算法, 引入 SOM 算法构建直觉模糊集, 使得分类器在多因素下综合判别无标记样本, 提高无标记样本的使用率, 从而在迭代中扩展有标记样本集. 在多个 UCI 数据上进行实验, 结果数据表明, 分类器的性能得到提高, 学习无标记样本过程是影响分类器的关键点.

关键词: Tri-Training 算法; SOM 算法; 直觉模糊集; 半监督

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2016)03-0134-04

The Improvement of Tri-Training Algorithm Based on Intuitionistic Fuzzy Sets

PENG Ya-qin¹, GONG Ning-sheng²

(1 Department of Computer Science and Engineering, Sanjiang University, Nanjing 210012, China;

2 Department of Electronic and Information Engineering, Nanjing University of Technology, Nanjing 210009, China)

Abstract: Tri-Training algorithm belongs to semi-supervised algorithm, the initial classifier performance of the algorithm is influenced by the labeled samples. When the number of the labeled samples is insufficient, the classifier performance is relatively weak, which will affect the iteration later. Therefore, an improvement algorithm called IFS-Tri-Training (Tri-Training based on intuitionistic fuzzy sets) is proposed, the SOM algorithm is applied in the paper to build intuitionistic fuzzy sets, so the classifier can measure the unlabeled samples by more factors, which makes the efficiency of the unlabeled samples is increased and labeled sample set is expanded. The experimental result on UCI datasets shows that the classifier performance is improved and the method of studying unlabeled samples is important.

Key words: Tri-Training algorithm; SOM algorithm; intuitionistic fuzzy sets; semi-supervised

1 引言

在机器学习中, 有标记的样本只占少数, 大多数都是无标记样本, 所以半监督学习应运而生, 它是一种介于监督学习和无监督学习之间的学习方式, 是机器学习的研究热点, 主要分为半监督分类和半监督聚类两大类^[1]: 半监督分类以原有的分类算法为基础, 把大量的无标记样本加入学习过程, 充分应用

了样本的信息, 从而提高分类的准确性; 半监督聚类是以聚类为基础, 加入少量标记样本, 辅助和修正聚类过程, 从而提高聚类效果. 半监督学习算法目前已经取得很多研究成果, 并且成功应用到网络检测、图像处理以及视频标注等领域.

半监督分类的算法很多: 文献[2]提出应用 EM 算法, 它把未标记样本属于某类别的概率当作一组隐含参数来输入, 然后应用算法进行标记和评估; 而

后有学者提出基于图的半监督分类算法,比如 Belkin 等人提出的流形正则化框架思想^[3],扩展了半监督算法;向量机也早就应用到半监督学习中,Joachims T^[4]提出了经典直推式支持向量机,算法采用局部的组合搜索方式;协同训练算法也比较流行,这类算法采用多个分类器,比如 Tri-Training 算法^[5]. 本文主要研究 Tri-Training 算法,该算法由 Zhou 和 Li 在 2005 年提出,进一步放松了协同训练算法的执行条件,算法不需要交叉验证,使用范围较广,而且效率很高. 算法中的初始分类器由有标记样本训练而成,所以算法初始阶段如果没有足够的学习样本,初始分类器就会相对弱化;而且在实际应用中,一般初始有标记数据都比较有限,更多存在的是无标记的数据,在此情况下,性能较弱的初始分类器会直接影响着后期的迭代过程,从而影响分类精度. 为此本文提出 IFS-Tri-Training 算法,在算法的迭代初始就引入神经网络中 SOM 算法,应用直觉模糊集进一步有效挖掘和扩展可利用的无标记样本数目,加强分类器的学习,从而提高分类器的性能.

2 Tri-Training 算法

半监督学习之中的一个常用技术是自学习,是指在学的过程中,首先用带标记的样本训练初始分类器,而后用这个分类器对无标记样本进行分类,并标记其类别号,加入到当前的样本集,并用扩展后的样本集重新训练分类器.

Tri-Training 算法的自学习过程引入了投票器机制,设置了三个分类器:一个主分类器,两个为辅助分类器. 学习过程中,首先应用 bootstrap 算法构建有标记样本数据集 L 来训练初始分类器;而后两个辅助分类器对无标记样本集 U 进行分类,标记结果相同的无标记样本被认为是置信度最高的样本扩充到样本集中,继续用于训练主分类器;最后当三个分类器都不再变化时,迭代结束. 算法中每一轮被挑选出来的无标记样本只参加本轮的学习,下一轮仍然作为无标记样本,避免了噪声数据过早被引入,而且作者证明了只要新标记的样本足够多,就足以抵消噪声带来的误差. 其算法流程如图 1 所示.

Tri-Training 算法跟其他的协同算法对比,对分类器类型没有约束,对数据集也没有特别要求,而且耗时较少,是一种广泛使用的协同训练算法.

但是在实际应用中,初始有标记数据集规模很小,不足以训练出高精度分类器^[6],当初始分类器能力较弱时,主分类器的性能直接影响其后的学习

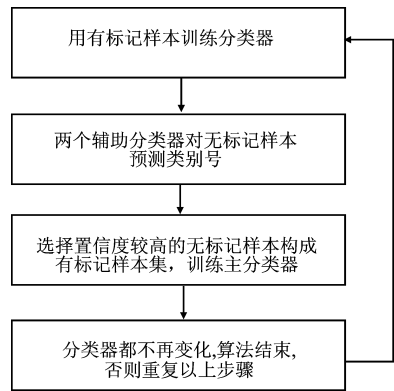


图1 Tri-Training 算法流程图

迭代过程. 所以在有标记样本较少的情况下,如何更加充分地应用无标记数据,一直是半监督学习需要解决的问题,对应到 Tri-Training 算法中如何更加充分地应用无标记样本也正是本文的研究重点.

3 IFS-Tri-Training 算法

在 Tri-Training 算法中,两个辅助分类器预测结果相同的无标记样本被认为置信度最高,并构成新的有标记样本集 L_i 参与下一次的迭代. 通过实验可以发现,当有标记样本不足时, $(U - L_i)$ 数据集中还有可利用的样本数. 分析其学习结果表现为两种情况:第一种是三个分类器的分类结果都不相同,第二种是主分类器的分类结果只与其中的一个辅助分类器分类结果相同,说明这类样本的分布特征有别于有标记样本的分布特征,有标记样本对这类样本的分布指导比较受限,因此可以考虑从无标记样本自身的分布特征出发,引入具备自组织能力的算法来综合判定 $(U - L_i)$ 数据集中无标记样本的类别,将这类样本加入到分类器的学习中. 为此本文构建了 IFS-Tri-Training 算法,引入自组织 SOM 算法,应用其对不同样本特征产生不同响应的特点,对 $(U - L_i)$ 数据集进行学习;同时为了更加准确地学习 $(U - L_i)$ 数据,提出结合直觉模糊集多因素下综合标注无标记样本. 挖掘这类样本能够提高无标记样本的使用率,并且补充反映样本的分布情况,可利用的样本越多,分类器的学习就越全面.

3.1 SOM 算法

SOM (Self-organizing Maps) 神经网络,由芬兰学者 Teuvo Kohonen^[7]提出,已被广泛地应用于数据挖掘、语音识别、图像识别等多个领域.

SOM 神经网络由输入层和输出层组成:输入层模拟视觉感官用于接收样本,输出层模拟大脑神经元,对输入样本产生不同响应,从而形成能够反映样

本分布情况的特征图.其基本算法如下所示.

(1) 输入样本 X_i^p , 计算 X_i^p 与神经元权值 ω_{ji} 之间的欧氏距离, 标注出欧式距离最小的获胜神经元.

(2) 定义优胜邻域 $N_j(t)$, 初始邻域 $N_j(0)$ 较大, 然后随着时间增大 $N_j(t)$ 逐渐收缩; 定义学习率 $\eta(t)$, 也是随着时间的增大而减小.

(3) 调整神经元的权值, 对优胜邻域 $N_j(t)$ 内的所有神经元按照式(1)修改权值:

$$\omega_{ji}(t+1) = \omega_{ji}(t) + \eta(t)(X_i^p - \omega_{ji}(t)) \quad i=1, 2, \dots, n, j \in N_j(t) \quad (1)$$

(4) 输入下一个样本, 回到步骤(1), 一直到所有样本都学习结束.

(5) $t = t + 1$, 返回步骤(1), 直至 t 达到学习次数 T 为止.

SOM 算法通过神经元竞争, 可以充分挖掘 $(U - L_i)$ 这类无标记样本自身的分布特征, 从而标注样本的类别.

3.2 直觉模糊集

直觉模糊集^[8]是模糊集的一种成功推广^[9], 其在模糊系统理论的发展过程中另辟蹊径, 提出了犹豫度等概念, 能够比较灵活地描述不确定性, 已经被广泛应用.

定义1 若 X 上2个映射 $\mu_{\tilde{A}}: X \rightarrow [0, 1]$ 和 $\nu_{\tilde{A}}: X \rightarrow [0, 1]$, 若存在 $x \in X \mid \mu_{\tilde{A}}(x) \in [0, 1]$ 和 $x \in X \mid \nu_{\tilde{A}}(x) \in [0, 1]$, 并满足条件 $0 \leq \mu_{\tilde{A}}(x) + \nu_{\tilde{A}}(x) \leq 1$, 则称 $\mu_{\tilde{A}}(x)$ 和 $\nu_{\tilde{A}}(x)$ 确定了论域 X 上的一个直觉模糊集 $\tilde{A}$, 可简记为 $\tilde{A} = \{ \langle x, \mu_{\tilde{A}}(x), \nu_{\tilde{A}}(x) \rangle \mid x \in X \}$, 分别称 $\mu_{\tilde{A}}(x)$ 为元素 x 属于 $\tilde{A}$ 的隶属度, $\nu_{\tilde{A}}(x)$ 为元素 x 属于 $\tilde{A}$ 的非隶属度, $\pi_{\tilde{A}}(x) = 1 - \mu_{\tilde{A}}(x) - \nu_{\tilde{A}}(x)$ 为 $\tilde{A}$ 中元素 x 的犹豫度. 对应到算法中, $\mu_{\tilde{A}}(x)$ 是样本属于某个类别的隶属度, $\nu_{\tilde{A}}(x)$ 是样本属于某个类别的非隶属度, $\pi_{\tilde{A}}(x)$ 是分类器没有对该样本进行有效分类, 不在类别域.

定义2 设 $A_j = \langle \mu_j, \nu_j \rangle (j = 1, 2, \dots, n)$ 是直觉模糊集. 若映射 $f_w^A: F^n \rightarrow F$ 使得

$$f_w^A(A_1, A_2, \dots, A_n) = \sum_{j=1}^n \omega_j A_j \quad (2)$$

则称 f_w^A 为直觉模糊集加权集结算子.

综合定义1和定义2可推出定理1

定理1 设 $A_j = \langle \mu_j, \nu_j \rangle (j = 1, 2, \dots, n)$ 是直觉模糊集, 则由直觉模糊集加权集结算子运算得到的结果仍为直觉模糊集, 且

$$f_w^A(A_1, A_2, \dots, A_n) = \langle 1 - \prod_{j=1}^n (1 - \mu_j)^{\omega_j}, \prod_{j=1}^n \nu_j^{\omega_j} \rangle \quad (3)$$

根据定理1可知, 如果存在多个因素下关于 x 样本类别判定的直觉模糊集, 就可以通过加权计算得出其综合评判值.

3.3 构建 IFS-Tri-Training 算法

3.3.1 构建直觉模糊集

IFS-Tri-Training 算法中应用直觉模糊集综合求解样本的类别是一个应用多决策者意见的判断过程, 此方法最关键的就是建立评价指标体系. 在 SOM 算法中, 初始权值和样本集是算法的核心影响因素, 所以可以设置直觉模糊集模型如下所示.

$A_{\text{sample}}^{W-SOM}$ 表示在 SOM 算法权值因素下关于样本类别的直觉模糊集. 初始权值决定着 SOM 算法的学习结果, 所以 n 个不同初始权值的 SOM 算法可表示为投票专家们, 标记每个专家在某类别下的投票数即可得出直觉模糊集.

$A_{\text{sample}}^{S-SOM}$ 表示在 SOM 算法样本因素下关于样本类别的直觉模糊集, 不同的学习样本集对 SOM 算法意味着不同知识集, 所以也可以对应为若干个专家参与学习投票.

根据投票结果, 即可按式(4)求出 $A_{\text{sample}}^{W-SOM}$ 和 $A_{\text{sample}}^{S-SOM}$ 这两个直觉模糊集.

$$\begin{aligned} \mu(x) &= \left(\sum_{i=1}^n \text{AGR}_i(c) \right) / n \\ \nu(x) &= \left(\sum_{i=1}^n \text{UAGR}_i(c) \right) / n \\ \pi(x) &= \left(\sum_{i=1}^n \text{ABST}_i(c) \right) / n \end{aligned} \quad (4)$$

式中, c 为样本类别, $\text{AGR}_i(c) (i \in 1, 2, \dots, n)$ 为赞成票, $\text{UAGR}_i(c) (i \in 1, 2, \dots, n)$ 为反对票, $\text{ABST}_i(c) (i \in 1, 2, \dots, n)$ 为弃权票.

通过上述设置, 就可以建立多个因素下关于 x 样本类别判定的直觉模糊集, 然后根据实验要求设置权重向量 ω , 按式(3)加权计算公式, 即可综合得出样本的类别.

3.3.2 IFS-Tri-Training 算法的步骤

根据以上的概念, IFS-Tri-Training 算法的基本步骤如下所示.

(1) 应用 $L_i \leftarrow \text{Bootstrap}(L)$ 对每个分类器 h_i 训练, 设置初始误差参数 $e'_i \leftarrow 0.5, l'_i \leftarrow 0$.

(2) $L_i \leftarrow \varphi$, 计算 $e_i = \text{MeasureError}(h_j \& h_k) (j, k \neq i)$, 如果 $e_i < e'_i$, 则对 U 中每个样本运用辅助分类器预测类别号, 如果两个辅助分类器预测标号相同, 则把此样本加入 L_i 中.

(3) 构建 $A_{\text{sample}}^{W-\text{SOM}}$. 将 $(U - L_i)$ 样本输入到 n 个 SOM 网络, 设置不同的初始化权值, 邻域、学习率等参数一致, 通过网络自适应学习和参数调整, 最后得出获胜单元, 并标记出样本的类别, 根据 (4) 得出 $A_{\text{sample}}^{W-\text{SOM}}$.

(4) 构建 $A_{\text{sample}}^{S-\text{SOM}}$. 生成若干个 SOM 网络, 保证初始参数一致, 输入对应的样本集 $S \supseteq (U - L_i)$ 并学习, 标记分类结果根据 (4) 得出 $A_{\text{sample}}^{S-\text{SOM}}$.

(5) 设置权重向量 ω , 根据式 (3), 计算出样本的综合评价结果, 如果某类别的结果大于平均投票将此样本扩展到 L_i 中.

(6) 如果 $l'_i = 0$, 则计算新的 $l'_i \leftarrow [\frac{e_i}{e'_i - e_i} + 1]$; 如果 $l'_i < |L_i|$, 而且如果 $e_i |L_i| < e'_i l'_i$, 则 $\text{update_}i \leftarrow \text{true}$; 否则如果 $l'_i > \frac{e_i}{e'_i - e_i}$, 从样本集 L_i 随机抽取出 $[\frac{e'_i l'_i}{e'_i} - 1]$ 样本放入 L_i 中, 最后 $\text{update_}i \leftarrow \text{true}$.

(7) 如果每个 $\text{update_}i \leftarrow \text{true}$, 则将 L 和 L_i 合并, 训练分类器, 然后设置 $e'_i \leftarrow e_i, l'_i \leftarrow |L_i|$.

(8) 进入下一轮迭代, 如果三个分类器 h_i 都没有变化, 迭代结束.

通过上述的改进后, 进一步对 $(U - L_i)$ 数据集中无标记样本进行了标注, 提高了无标记样本的使用率, 扩展了有标记样本. 虽然有可能会错误标注样本, 但算法已证明当样本的数目较多时, 可以抵消误差.

4 实验分析

本文的实验数据采用 UCI 机器学习库中的经典数据集: Iris、Wine 和 Ionosphere, 数据集样本信息如表 1 所示.

表 1 实验数据集

数据集	样本集	属性数	类别数
Iris	150	4	3
Wine	178	13	3
Ionosphere	351	34	2

IFS-Tri-Training 算法的分类器采用决策树算法, $A_{\text{sample}}^{W-\text{SOM}}$ 下设置三个 SOM 算法专家, 其初始权值由随

机函数生成, 并归一化操作; $A_{\text{sample}}^{S-\text{SOM}}$ 下也是三个 SOM 算法专家, 其样本集为 $L + U, U, (U - L_i)$; 由于样本的因素对网络的影响大于初始权值的因素, 所以权值向量设为 $\omega = \{0.4, 0.6\}$. 实验选择 50% 的样本作为训练样本, 其中无标记样本占 70%, 剩下的 50% 的样本作为测试样本, 在相同的实验条件下进行了 10 次独立的半监督学习实验, 并以 10 次独立实验的平均错分率作为算法的衡量指标. 表 2 中分别列出了三个数据集的两种算法下的平均错分率.

表 2 分类实验结果

数据集	Tri-Training	IFS-Tri-Training
	算法错分率	算法错分率
Iris	5.47	4.13
Wine	21.01	13.59
Ionosphere	18.58	13.07

对比实验结果, 改进后的 IFS-Tri-Training 算法平均错误率有所下降, 由此可见, 增加的无标记样本扩展了总体的样布分布, 分类器的分类能力得到了提高, 图 2 中对比了两种算法下的未标记样本使用率.

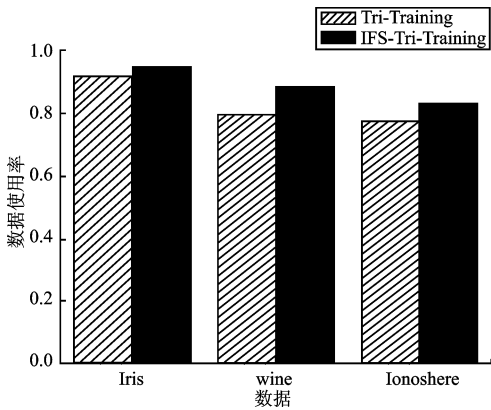


图 2 两种算法的未标记样本使用率对比图

5 结束语

半监督学习能够有效利用有标记样本和无标记样本, 从而提高学习效果, 其中的 Tri-Training 算法更是广泛使用. 本文提出的 IFS-Tri-Training 算法, 应用直觉模糊集通过 SOM 算法挖掘更多的无标记样本. 实验结果表明, 改进算法的分类效果有了一定的提高, 新算法确实能够提高无标记样本的使用率, 使得总体有标记样本的信息量增大. 下一步的工作是应用直觉模糊集去综合判定所有的无标记样本, 提高样本的标注正确率, 从而加强分类器的学习.

(下转第 141 页)

差为

$$\xi = \frac{0.021 \text{ MPa}}{60 \text{ MPa}} \times 100\% = 0.035\%.$$

以上结果说明采用结合二次多项式拟合和三次样条插值算法融合电压信息和温度信息后,系统的性能得到显著提升,较好地消除了温度对压力检测结果的影响。

6 结束语

依据算法原理及原始标定数据提出结合曲线拟合和三次样条插值的信息融合算法. 首先固定温度不变,对数据点进行二次多项式拟合,得到压力和电压的关系曲线,再对拟合获得的压力值与标准温度进行三次样条插值,完成压力和温度信息融合,最终得到经过信息融合消除温度影响的压力值,以显示正确的压力结果. 实验验证了该算法的有效性,这样在标定点较少的情形下使压力检测系统的温度补偿达到较高的精度。

参考文献:

- [1] Pramanik. Temperature compensation of piezoresistive micro-machined porous silicon pressure sensor by ANN [J]. Microelectronics Reliability, 2006, 46 (2): 343-351.
- [2] Kowalski. Miniature pressure sensors and their temperature compensation [J]. Sensors and Actuators, 1987, 11 (4): 367-376.
- [3] 林成森. 数值计算方法 [M]. 北京: 科学出版社, 1998.
- [4] 孙志忠, 袁慰平, 闻震初. 数值分析 [M]. 南京: 东南大学出版社, 2011.
- [5] 王宝君, 刘欣菊, 刘宝琦. 压力校准技术及应用介绍 [J]. 计测技术, 2008, 28(3): 64-65.
- [6] 刘庆, 王昌. 基于 Vague-Topsis 法的多传感器信息融合能效评价研究 [J]. 微电子学与计算机, 2014, 31 (11): 51-55.

作者简介:

夏菽兰 女, (1969-), 硕士, 副教授. 研究方向为电子技术与信息处理. E-mail: xslnj@126.com.

赵力 男, (1958-), 博士, 教授, 博士生导师. 研究方向为信号与信息处理.

(上接第 137 页)

参考文献:

- [1] Olivier C, Bernhard S, ALEXANDER Z. Semi-supervised learning [M]. Cambridge: MIT Press, 2006.
- [2] Nigam K, McCallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using EM [J]. Machine learning, 2000, 39(2-3): 103-134.
- [3] Belkin M, Niyogi P, Sindhwani V. A geometric framework for learning from labeled and unlabeled examples [J]. Journal of machine learning research, 2006, 7 (11): 2399-2434.
- [4] Joachims T. Transductive inference for text classification using support vector machines [C] // Proc of the 16th Int'l Conf on Machine Learning (ICML'99). San Francisco: Morgan Kaufmann Publishers, 1999: 200-209.
- [5] Zhou ZH, Li M. Tri-training: exploiting unlabeled data

using three classifiers [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.

- [6] 邓超, 郭茂祖. 基于 Tri-Training 和数据剪辑的半监督聚类算法 [J]. 软件学报, 2008, 19(3): 663-673.
- [7] Kohonen T. The Self-organizing Maps [J]. Proceedings of the IEEE, 1990, 78(9): 1464-1480.
- [8] Atanassov K T. Intuitionistic fuzzy sets [J]. Fuzzy Sets and Systems, 1986, 20(1): 87-96.
- [9] Atanassov K T. My personal view on intuitionistic fuzzy sets theory [J]. Studies in Fuzziness and Soft Computing, 2008, 14(220): 23-43.

作者简介:

彭雅琴 女, (1983-), 硕士, 讲师. 研究方向为神经网络. E-mail: pyqaya@163.com.

宫宁生 男, (1958-), 硕士, 教授. 研究方向为神经网络.