一种自适应的 Tri-Training 半监督算法①

彭雅琴[1], 宫宁生[2]

[1](三江学院 计算机科学与工程学院, 南京 210012)
[2](南京工业大学 计算机科学与工程学院, 南京 210009)

摘 要: Tri-Training 算法是半监督算法的一种, 在学习过程中容易错误标注无标记样本, 从而降低分类性能, 为此提出一种 ADP-Tri-Training(Adaptive Tri-Training)算法, 改进协同工作方式, 根据几何中心设置分类器组成, 然后应用模糊数学理论将多个独立的分类器组合, 使得算法可以在多因素下综合评价样本, 并在此基础上引入遗传算法动态设置组合权重以适应于具体的样本集, 从而尽可能降低样本标注的错误率, 多个实验结果表明 ADP-Tri-Training 算法具有更好的分类性能.

关键词: Tri-Training 算法;自适应;遗传算法;差异性度量;半监督

Adaptive Tri-Training Semi-Supervised Algorithm

PENG Ya-Qin[1], GONG Ning-Sheng[2]

[1](Department of Computer Science and Engineering, Sanjiang University, Nanjing 210012, China)
[2](Department of Computer Science and Engineering, Nanjing University of Technology, Nanjing 210009, China)

Abstract: Tri-Training algorithm belongs to semi-supervised algorithm,unlabeled samples are often labeled incorrectly in study, and the performance is affected. So the ADP-Tri-Training (Adaptive Tri-Training) algorithm is proposed, cooperative work mode is improved, a classification composition scheme based on geometric center is used, the fuzzy mathematics theory is applied to combine the classifiers, so the algorithm can evaluate the samples by multiple factors, genetic algorithm is introduced to dynamically set the combined weight in order to adapt different sample sets, also it can reduce the error of classifies as far as possible, finally the experimental results show that the proposed algorithm is more effective.

Key words: Tri-Training algorithm; adaptive; genetic algorithm; novel diversity measure; semi-supervised

在实际应用中, 有标记样本获取相对困难, 更多的是无标记样本, 因此半监督算法不断被提出, 近年来取得了飞速的发展, 并广泛应用于图像识别、入侵检测、文档识别等领域. 半监督算法介于有监督算法和无监督算法之间, 有半监督聚类和半监督分类两种, 本文研究的是半监督分类算法, 经典的算法有: 基于 EM 算法的半监督模型[1], 其把无标记样本属于某个类别的概率当作隐含参数输入, 然后应用 EM 算法进行评估. 支持向量机的半监督方法也被提出, 基本原理是利用有标记样本和无标记样本寻找最大边缘的超平面, 使得各个样本之间间隔最大[2]. 基于图的半监督学习思想将图中结点对应为样本, 边对应为相似度, 然后进行类别标记, 比如 Belkin 等人提出的流形正则化框架思想[3]. 协同训练算法的根本是使用多个分类器, 将置信度较高的无标记样本数据加入到分类器的训练中,比如 Tri-Training 算法[4], 由 Zhou 和 Li 提出, 其框架对样本和分类器没有任何约束, 既利用了多分类器的协同优势又避免了传统协同策略验证时间长, 效率很高. 但是在该算法中, 无标记样本容易被错误地标注并积累而影响学习性能的提高, 究其根本原因还是多个分类器协同工作机制较弱, 为此本文提出 ADP-Tri-Training (Adaptive Tri-Training) 算法, 调整分

① 收稿时间:2015-11-30;收到修改稿时间:2016-01-18 [doi:10.15888/j.cnki.csa.005298]

类器的设置规则, 然后引入模糊数学将多个独立分类器按权重组合, 并且引入遗传算法动态求解权重向量, 从而加强无标记样本的标注能力, 且自适应于不同的样本集, 分类器的总体性能得到提高.

1 Tri-Training算法介绍

Tri-Training 算法的基本思路是多分类器思想, 设置了三个分类器 C1, C2, C3, 然后应用 bootstrap 算法随机采样有标记样本集 LL 形成有差异的训练数据集, 并训练三个分类器, 使得三个分类器具备不同的分类能力. 然后三个分类器对无标记样本集 UL 进行分类, 如果 C2 和 C3 对 X 的分类结果一样, 则将 X 样本的类别号设为 C2(X)并加入到 C1 的训练集中, 其他两个分类器同理类推. 多次迭代以后, 当三个分类器的分类性能都没有变化时训练停止. 为了约束噪声数据的引入, 定义了误差约束公式(1):

$$|\boldsymbol{LL}\cup\boldsymbol{L}^{l}|(1-2\frac{\eta_L|\boldsymbol{LL}|+e^{l}|\boldsymbol{L}^{l}|}{|\boldsymbol{LL}\cup\boldsymbol{L}^{l}|})^2<|\boldsymbol{LL}\cup\boldsymbol{L}^{l+1}|(1-2\frac{\eta_L|\boldsymbol{LL}|+e^{l+1}|\boldsymbol{L}^{l+1}|}{|\boldsymbol{LL}\cup\boldsymbol{L}^{l+1}|})^2 \quad (1)$$

η_L 有标记样本的分类噪声率, e^l 为第 l 轮中非主分类器的分类误差率的上限, $\boldsymbol{L}^l$ 为第 l 轮迭代中非主分类器的对无标记样本中分类标记相同的样本集.

2 ADP-Tri-Training算法原理和步骤

Tri-Training算法实际上属于多分类器系统, 其中分类器对算法的影响甚大. 生成分类器的方法有很多, 例如选取不同的分类模型或参数, 不同的特征子空间或训练样本等[5]. Tri-Training算法中分类器采用了相同的分类算法, 然后重采样有标记样本构造分类器的差异性, 但是在有些情况下有标记样本较少, 不足以训练出较大差异的分类器, 因此标注样本投票时, 会趋于一致. 而且在投票无标记样本时, 采用的是平均投票机制, 没有考虑到分类器自身的强弱, 这都可能降低无标记样本的正确使用率. 因此本文提出改进思想, 采用新的分类器设置方法, 而后引入模糊数学理论设置评价因素、评价权重组合分类器, 并且应用遗传算法动态设置分类器的权重, 使得算法中的分类器尽可能独立, 且能够根据每个分类器的能力强弱为无标记样本投票并标注, 从而加强算法的学习过程.

2.1 分类器设置

多分类器系统的性能很大程度上取决于成员分类器的差异性[6], 因此 ADP-Tri-Training 算法中首要解决的问题就是分类器的生成, 其必须满足以下基本要求:

① 为了提高分类器的独立性, 要求每个分类器采用的算法各不相同: 比如 BP 算法、贝叶斯算法这类单分类器, 或者本身就是多分类器的 AdaBoost 算法等都可以.

② 分类器的性能不能太差, 否则集成的分类器的精确度也不会高.

③ 分类器的结果应该具备多样性. 若每个分类器对错误样本的标注都是一样的, 那么其对集成分类器性能的提高没有作用. 所以本文采用了基于几何中心的差异性度量方法[5], 由梁绍一等提出, 其基本原理是将样本映射到圆周中, 然后计算几何中心用来表示分类器之间的差异性. 分类器的几何中心越分散, 分类器之间的差异就越大. 该方法能有效的表示分类器之间的差异性, 而且能够很好地克服“差异性淹没”问题. 方法的具体步骤为:

1) 假设样本类别为 $1\sim CL$, 将圆周做 CL 个等分, 将每个类别的中心点 Cen_{cl} 按照先后顺序放在等分的圆弧中心.

2) 如果样本被分类器正确分类, 记为有效样本 $\boldsymbol{X}$. 假设其类别为 cl, 则

$$PreC=\begin{cases}cl-1 & (cl\neq 1)\\ CL & (cl=1)\end{cases} \quad (2)$$

$$PosC=\begin{cases}cl+1 & (cl\neq CL)\\ 1 & (cl=CL)\end{cases} \quad (3)$$

计算样本 X 与 Cen_{PreC}、Cen_{PosC} 之间的欧式距离分别为 $d_{X,PreC}$、$d_{X,PosC}$.

3)将样本点 $\boldsymbol{X}$ 映射到弧 $\overset{\frown}{Cen_{PreC}Cen_{cl}Cen_{PosC}}$ (Cen_{PreC} 顺时针至 Cen_{PosC})上, MP_X 为样本在圆周上的映射点, 那么其应该满足关系

$$\frac{AL(Cen_{PreC},MP_X)}{AL(MP_X,Cen_{PosC})}=\frac{d_{x,Prec}}{d_{x,Posc}} \quad (4)$$

其中 $AL(a,b)$ 表示弧 $\overset{\frown}{ab}$ (a 顺时针至 b)的长度.

4) 按照(2)和(3)对所有的有效样本进行映射, 而后计算多个分类器下所有映射点的几何中心 $Z_j(\mathrm{j}=1,2,\cdots n)$.

5) 多分类器集合 C_n 的差异性为

$$Diversity(\mathrm{C}_n)=\sum_{j=1}^{n}d(\mathrm{Z}_j,\mathrm{Z}_{ave}) \quad (5)$$

其中 Z_{ave} 为所有分类器所得中心的平均位置.

2.2 分类器组合

多分类器组合方法已经成为机器学习和模式识别

的重要分支[7], 研究成果颇多. 所谓多分类器组合是根据各个分类器的整体性能将每个分类器的输出按某种方式“组合”到一起, 并达到共识[8], 所以“组合”实际上是为了能够在多个因素下综合评价样本, 因此这个过程可以理解如下: 对样本建立评价因素、评价权重、评价结果等参数, 然后采用运算规则进行综合评价样本, 即为模糊综合评价方法, 因此设定评价模型如下:

(1) 评价因素 U

通过多种角度对目标对象进行评价才全面且合理, 所以构建评价因素非常有必要. 多分类器下, 通常是将不同分类器的结果进行融合, 以提高分类识别效果及鲁棒性[9]. 因此在模糊数学中的多个评价因素就可以设定为 3 个分类器, 且已经证明, 分类器的独立性越强, 综合判定的准确率就越高.

(2) 评价结果 V

评价结果由每个分类器的类别标记决定, 结果域为属于和不属于, 用于表示样本是否属于某类别的概率.

(3) 评价权重 W

权重主要是体现因素的重要性, 常见的权重设置方法有: 专家投票、层次分析法、熵值法等. 分类器权重的设定是分类器集成方法中的重中之重, 直接影响着集成分类器的性能, 考虑到静态的设定权重不一定对每个数据集都适用, 所以本文提出结合遗传算法动态的得出每个分类器的权重, 以适应于不同的数据集, 且获得较高的分类性能.

2.3 遗传算法设置权重

遗传算法是由美国 John Holland[10]教授最早提出, 是一种基于生物进化论中的“适者生存、优胜劣汰”进制的优化方法, 具有较强的全局搜索能力, 且与问题领域无关, 过程简单. 算法中把待求解问题转化为染色体, 通过循环进化过程, 最终可以得到求解问题的满意解.

其基本过程是: 首先根据求解问题对染色体编码, 产生一定数目的种群 m, 并计算每个种群的适应度函数 $f_j(j\in m)$, 然后根据一定的原则对种群进行选择, 并进行交叉和变异, 产生子代. 这一过程重复进行, 直到满足优化准则为止.

算法的核心参数有:

(1) 权重编码

每个分类器在投票时, 所发挥的作用不一样, 所以有必要设置权重将分类器有效的组合起来. 所以组合权重就是求解问题, 必须对其先进行编码, 解码后还需通过公式(6)转化权重, 使得组合权重之和转化为 1.

$$\mathrm{W}_i^{'} = W_i / \sum_{i=1}^{n} W_i \tag{6}$$

(2) 设计适应度

适应度是用来衡量种群的优劣程度, 非常重要, 对应到算法中就是该权重向量下集成分类器的性能越高越好. 所以算法中将设置小部分有标记样本作为验证样本集 $\boldsymbol{Y}$, 记初始分类器对验证样本集的分类精度为 ε_j, 学习过程结束时最终分类器对验证样本集的分类精度为 $\varepsilon_j^{'}$, 则按照公式(7)可得出种群的适应度.

$$f_j = \varepsilon_j^{'} - \varepsilon_j \tag{7}$$

适应度越高, 说明无标记样本的学习就越有效, 分类器的本次学习能力提高的也就更多.

(3) 其他算子

选择算子, 使用轮盘选择算子, 公式如(8); 交叉算子, 采用单点交叉; 变异算子, 采用基本位变异.

$$s(j) = f_j / (\sum_{j=1}^{m} f_j) \tag{8}$$

2.4 ADP-Tri-Training 算法步骤

综合以上分析, ADP-Tri-Training的基本步骤如下所示:

1) 按照算法中的分类器设置规则选择3个独立的分类器 $\mathrm{C}(i)(i\in 1,2,3)$ 即为评价因素 $\boldsymbol{U}$, 设置初始误差参数 $e^l{}_i \leftarrow 0.5$, 初始无标记学习样本规模 $|\boldsymbol{L}_i^l| \leftarrow 0$; 设置种群个数, 然后进行染色体编码.

2) 在有监督方法下应用 $\boldsymbol{LL}$ 训练每个分类器 $\mathrm{C}(i)$, 并记录错误率 $e_i^{l+1} = Error(\mathrm{C}(j)\,\&\,\mathrm{C}(k))(j,k\neq i)$.

3) 如果 $e^{l+1}{}_i < e^l{}_i$, 对无标记样本集 $\boldsymbol{UL}$ 进行学习, 根据专家投票对每个样本生成评价因素下的关于某个类别的评价结果 $\boldsymbol{V}$.

4) 染色体个体转化后赋值给权重向量 $\boldsymbol{W}$, 然后按照模糊运算规则计算综合评价结果, 如果综合评价结果超过平均值则生成样本类别, 标注无标记样本加入样本集 $\boldsymbol{L}_i^{l+1}$.

5) 当 $|\boldsymbol{L}_i^l|$ 为 0 时, $|\boldsymbol{L}_i^l| \leftarrow \left\lfloor \frac{e_i^{l+1}}{e_i^l - e_i^{l+1}} + 1 \right\rfloor$; 如果 $|\boldsymbol{L}_i^l| \lessdot |\boldsymbol{L}_i^{l+1}|$, 且 $e_i^{l+1}|\boldsymbol{L}_i^{l+1}| < e^l{}_i|\boldsymbol{L}_i^l|$, 则 $flag\_i \leftarrow true$; 否则如果 $|\boldsymbol{L}_i^l| > \frac{e_i^{l+1}}{e^l{}_i - e_i^{l+1}}$, 从 $\boldsymbol{L}_i^{l+1}$ 中抽取若干样本构成新的

$\boldsymbol{L}_i^{l+1}$, 最后 $flag\_i \leftarrow true$.

6) 对每个 $flag\_i \leftarrow true$, 将 $\boldsymbol{LL}$ 和 $\boldsymbol{L}_i^{l+1}$ 合并, 训练分类器 C(i), 然后设置 $e^l{}_i \leftarrow e_i^{l+1}$, $|\boldsymbol{L}_i^l| \leftarrow |\boldsymbol{L}_i^{l+1}|$.

7) 重复2)-6), 一直到所有的分类器性能稳定为止.

8) 对每个种群进行2)-7)上述操作, 而后分别引入验证样本集 $\boldsymbol{Y}$, 计算此样本集下的适应度函数 f_i, 然后采用轮盘选择算子选择适应度较高的种群, 按照一定的交叉方法和变异方法生成新的个体.

9) 重复 2)-8)的操作, 一直到满足优化准则为止, 选择最高适应度的组合分类器, 构成最终分类器模型.

3 实验

本文选用了 UCI 数据集[11]中的 6 个数据进行了实验, 如表 1 所示.

表 1 实验数据

数据集	属性	数目	类别
australia	14	690	2
germany	24	1000	2
wine	13	178	3
diabetes	8	768	2
dermatology	34	366	6
optdigits	64	5620	10

随机选择 70%的样本作为学习使用, 30%样本作为测试使用. 其中有标记样本与无标记样本的比例分别为 1:5 和 1:3, 再从有标记样本中抽取 20%作为验证样本集, 每组实验分别进行 10 次, 并将 10 次独立实验的错分率平均值作为最终实验结果. 分类器的设置根据样本而定, 本文假设候选分类器为 5 种算法: BP(Back Propagation)、J48、 Logistic、NB(Naive Bayes)和 AdaBoostM1(弱分类器为 Decision Stump). 遗传算法的参数为: 设置每个权重的值在[0,1]之间, 精度为小数点后 1 位即可, 二进制编码只需要 4 位, 交叉率为 0.4, 变异率 0.05, 种群规模 30, 迭代次数 40.

首先根据规则设置分类器, 应用 5 个候选分类器对每个数据集的有标记样本分类, 按照性能优劣程度选择其中 4 个分类器, 然后根据基于几何中心差异性度量公式计算任意组合的 3 个分类器差异性, 选择差异性最大的组合作为算法中的评价因素, 如表 2 所示.

表 2 评价因素

数据集	分类器 1	分类器 2	分类器 3
australia	AdaBoostM1	J48	BP
germany	AdaBoostM1	BP	Logistic
wine	BP	Logistic	NB
diabetes	AdaBoostM1	Logistic	BP
dermatology	NB	BP	Logistic
optdigits	BP	Logistic	NB

根据表 2 设置每个样本集的评价因素, 然后在每个数据集上分别应用 ADP-Tri-Training 算法和 Tri-Training 算法进行了对比实验, 表 3、表 4、表 5 分别列出了有标记样本和无标记样本比例为 1:5、1:3、1:1 的对比结果.

表 3 LL:UL=1: 5 对比实验结果

数据集	ADP-Tri-Training 错分率(%)	Tri-Training (分类器 1) 错分率(%)	Tri-Training (分类器 2) 错分率(%)	Tri-Training (分类器 3) 错分率(%)
australia	14.35	16.38	17.87	18.36
germany	27.33	29.23	30.03	30.1
wine	3.97	5.25	5.6	18.72
diabetes	22.47	24.74	24.96	28.83
dermatology	2.91	3.46	7.95	9.01
optdigits	2.86	3.61	7.83	9.79

表 4 LL:UL=1: 3 对比实验结果

数据集	ADP-Tri-Training 错分率(%)	Tri-Training (分类器 1) 错分率(%)	Tri-Training (分类器 2) 错分率(%)	Tri-Training (分类器 3) 错分率(%)
australia	15.12	17.97	18.36	18.07
germany	26.53	27.1	29.37	27.2
wine	2.83	3.02	5.85	18.3
diabetes	22.1	25.13	23.04	28.17
dermatology	2.01	2.42	3.83	4.98
optdigits	2.78	3.18	6.94	9.48

表 5 LL:UL=1: 1 对比实验结果

数据集	ADP-Tri-Training 错分率(%)	Tri-Training (分类器 1) 错分率(%)	Tri-Training (分类器 2) 错分率(%)	Tri-Training (分类器 3) 错分率(%)
australia	15.05	16.51	18.19	18.14
germany	26.3	26.9	28.57	26.8
wine	1.83	1.96.	4.85	16.36
diabetes	21.87	22.1	22.3	23.78`
dermatology	1.97	2.2	3.53	4.52
optdigits	2.32	2.61	5.34	8.4

通过对比的实验结果可得, 针对每个数据集 ADP-Tri-Training 算法的错分率都低于 Tri-Training 算法中最低错分率, 说明了 ADP-Tri-Training 算法的有

效性; 而且当有标记样本和无标记样本比例为 1:5 时, ADP-Tri-Training 算法的改进程度较好, 主要是在此比例下有标记样本集较小, 对无标记样本的学习指导能力较弱, 说明算法中的多因素下权重投票可以有效指导学习过程, 提高算法的识别性能.

4 结语

由于 Tri-Training 算法中无标记样本容易错误标注, 所以提出改进思路, 利用模糊数学理论将分类器集成, 并设置验证样本集作为适应度依据, 然后应用遗传算法根据样本动态设置分类器权重, 通过对比实验结果可得, ADP-Tri-Training 算法能够有效提高分类器的性能. 算法中验证样本集是个重要因素, 所占有标记样本的比率比较关键, 必须权衡考虑到有标记样本学习集; 同时验证样本集的错分率也会影响分类精度, 一般来说, 初始分类器对验证样本集的错分率越低越好. 遗传算法中的组合权值只是满意解, 在今后的研究过程中, 将进一步尝试其他适应度判别公式, 从而获得更优解.

参考文献

1 Nigam K, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Machine Learning, 2000, 39(2-3): 103–134.

2 Joachims T. Transductive inference for text classification using support vector machines. Proc. of the 16th Int'l Conference on Machine Learning (ICML'99). 1999. 200–209.

3 Belkin M, Niyogi P, Sindhwani V. A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Reseach, 2006, 7(11): 2399–2434.

4 Zhou ZH, Li M. Tri-Training: exploiting unlabeled data using three classifiers. IEEE Trans. on Knowledge and Data Engineering, 2005, 17(11): 1529–1541.

5 梁绍一,韩德强,韩崇昭.一种基于几何关系的多分类器差异性度量及其在多分类器系统构造中的应用.自动化学报, 2014,40(3):449–458.

6 李士进,常纯,余宇峰,王亚明.基于多分类器组合的高光谱图像波段选择方法.智能系统学报,2014,9(3):372–378.

7 Sun S. Local within-class accuracies for weighting Individual outputs in multiple classifier systems. Pattern Recognition Letters, 2010, 31(2): 119–124.

8 Zhou ZH. Ensemblemethods:foundations and algorithms. Boca Raton: CRC Press, 2012.

9 邢楠,朱虹,王栋,侯浩录.基于多分类器融合的图像真伪鉴别方法.计算机工程与应用,2014,50(24):164–167.

10 Holland JH. Adaptationnatural and artificial systems. 1st ed.Cambridge, MA:MITPress, l975: 12–35.

11 UCI repository of machine learning databases. http://www.ics.uci.edu/～mlearn/MLRository.html.