文章编号: 1001-0920(2011)11-1616-05

基于统计证据的半监督多分类器融合方法

孔志周[1,2], 蔡自兴[1]

(1. 中南大学 信息科学与工程学院，长沙 410083；湖南大学 金融与统计学院，长沙 410079)

摘 要: 针对半监督学习中未标记示例导致性能下降的问题, 提出一种新的协同训练算法LDL-tri-training. 首先通过最小显著性差异(LSD)假设检验方法使得3个成员分类器两两之间具有显著性差异; 然后采用D-S证据理论提高标注的稳定性; 最后利用局部异常因子检测算法剔除误标记的噪声样本. 实验表明, 与其他方法相比, LDL-tri-training算法具有较高的分类精度和稳定性.

关键词: 半监督；协同训练；最小显著性差异；统计证据；局部异常因子检测算法

中图分类号: TP301.6 **文献标识码:** A

Study on semi-supervised ensemble multiple classifiers based on statistical evidence

KONG Zhi-zhou[1,2], *CAI Zi-xing*[1]

(1. College of Information Science and Engineering，Central South University，Changsha 410083，China；2. College of Finance and Statistics，Hu'nan University，Changsha 410079，China. Correspondent：KONG Zhi-zhou，E-mail: zhizk@yahoo.com.cn)

Abstract: For the performance degradation of unlabeled data in semi-supervised learning, a new cooperative training algorithm, LDL-tri-training, is proposed. Firstly, by using least significant difference(LSD) hypothesis testing method, significant differences among three classifiers are tested. Then a D-S evidence theory is adopted to improve the stability of unlabeled data. Finally, local outlier factor(LOF) algorithm is used to reject error labeled data. Experiments show that LDL-tri-training can more effectively and stably utilize the unlabeled examples to improve classification generalization.

Key words: semi-supervised；cooperative training；least significant difference；statistical evidence；local outlier factor algorithm

1 引 言

一般而言, 要获取大量有标记的示例, 大都相对较为困难, 因为获得这些标记可能需要耗费大量的人力和物力. 多数情况下仅有少量的有标记示例, 因此如何充分利用这些资料进行半监督学习是一个非常关键的问题.

半监督的多分类器融合方法也称为“协同训练算法”, 此类算法隐含地利用了聚类假设或流形假设, 它们使用2个或多个学习器, 在学习过程中, 这些学习器进行相互标记, 从而使模型得以更新.

从目前协同训练算法的研究成果看, 所有这些算法中, 新产生的加标数据可以直接加入到带标集合中, 不需要人工参与, 但往往在加入正确的加标数据的同时, 也会给训练集带来一些噪音数据, 随着循环次数的增加, 不断积累的噪音数据会对后续分类器的训练带来严重影响, 甚至导致性能的急剧下降, 尤其对于原有标记样本非常稀少的情况更加突出. 即使是Zhou等人[1]提出的tri-training算法也与其他协同训练算法所遇到的问题一样, 由于迭代学习过程中无标记样例常被错误地标记并积累而损害学习性能的提高, 甚至会更严重. 因为在tri-training算法中, 若初始学习器比较弱, 则在评估过程中将给第3个学习器引入更严重的噪音, 从而导致更严重的后果. 周志华等人[2]认为, 随着训练的不断进行, 自动标记的示例中的噪音会不断积累, 其负作用会越来越大, 并建议利用数据剪辑技术来发现和处理这

收稿日期: 2010-07-14；**修回日期:** 2011-01-27.

基金项目: NSFC重大专项基金项目(90820302)；湖南大学青年基金项目(0723)；湖南大学中央高校基本科研业务费专项资金项目(54).

作者简介: 孔志周(1974−), 男, 讲师, 博士生, 从事信息融合、数据挖掘的研究；蔡自兴(1938−), 男, 教授, 博士生导师, 从事人工智能、机器人学等研究.

些噪音数据. 针对这个问题, 人们已经提出了一些新的思想. Li 等人[3]提出的 SETRED 算法就是在 Co-training 特例算法 Self-training 的迭代训练过程中引入特定的数据剪辑技术来过滤自标记样例中的噪声. 邓超等人[4]提出了 ADE-tri-training, 将基于最近邻规则的 Depuration 数据剪辑操作引入 tri-training 算法, 目的是识别每次迭代可能产生的误标记样例并移除. Wang 等人[5]在提出半监督回归方法 COREG 算法时, 提出了一个选择标记置信度最高的未标记示例的准则——标记置信度最高的未标记示例是在标记后与学习器的有标记训练集最一致的示例, 并认为也可以用于分类.

总之, 探究未标记示例导致性能下降的真正原因, 深入研究协同训练机制, 更加有效地利用标注数据的标签信息和未标注数据设计出性能更优的协同训练算法将是未来的研究重点[2].

本文在文献 [1-6] 的基础上, 提出一种新的单视图协同训练算法 LDL-tri-training. 首先通过最小显著性差异 (LSD) 假设检验方法, 使得 3 个成员分类器两两之间具有显著性差异; 然后采用 D-S 证据理论提高标注的稳定性; 最后利用局部异常因子检测 (LOF) 算法剔除误标记的噪声样本. 实验表明, LDL-tri-training 算法在无标记样本所占比例为 20% ∼ 80% 时均具有较高的分类精度和稳定性.

2　算法组成

2.1　基于假设检验的成员分类器生成

在半监督学习中, Wang 等人[3]认为, 成员分类器间应有较大的差异, 这样才可以通过协同训练来利用未标记示例提高学习性能. 然而, 产生于有监督学习的分类器差异性度量并不太适合半监督学习, 半监督学习的分类器由于有监督样本的缺乏而使得对差异性度量要求更加精准.

统计学中的配对 t-test 假设检验方法是一种衡量样本间方差差异的重要方法, 近几年已被应用于模型差异的评估[6]. 然而, 该方法只能用于两两配对比较, 无法实现 3 个及以上之间的两两比较. 在这种情况下, 可采用 Fisher 的最小显著性差异 (LSD) 方法, 该方法可以实现两两多重比较, 并判定哪些模型之间存在差异性[7].

设 x_1, x_2 和 x_3 为 3 个待检验模型, 则根据模型输出结果建立每 2 个模型均值之差的置信区间估计如下:

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\mathrm{MSE}\Big(\frac{1}{n_1}+\frac{1}{n_2}\Big)}. \tag{1}$$

式中: $t_{\alpha/2}$ 是自由度为 $n_T - k$ (n_T 为所有模型总样本数, k 为模型数) 时, 使 t 分布的上侧面积为 $\alpha/2$ 的 t 值; MSE 为方差的估计量 (即误差均方), MSE 不受零假设是否为真的影响, 始终是一个无偏估计.

判定方法是根据生成的置信区间是否包含数值零来决策. 若区间不包含数值零, 则可得出两模型存在差异的结论.

2.2　基于统计证据的标注

在半监督学习中, 当数据集样本量很少, 或取样过程存在明显的偏差时, 则可能无法正确地估计真实的数据分布, 从而无法形成正确的目标函数. 当目标函数不能由训练集的数据正确地表示时, 则在这个训练集上进行学习便不能得到满意的泛化能力.

为了解决上述问题, 可采用证据理论来提高未标记样本的选择效率. 作为信息融合的主要方法之一, 证据理论在多分类器融合方面的研究已经成为热点问题. Dymitr 等人[8]更明确提出了“由基于数据的方法向基于知识的方法转变将是未来的发展方向, 而证据理论是未来发展的 3 大热点方法之一”.

为更好地利用分类器的输出信息, 可采用度量层信息构造证据的基本概率分配函数. 假设有 R 个分类器对同一样本进行分类, 第 k 个分类器的输出为 xk, $k=1,2,\cdots,R$. 分类器融合的目的是将给定的样本划分到 M 个互不相交的模式类集合 $A_1, A_2, \cdots, A_M$ 中的一个.

假设给定一个样本, 可定义第 k 个分类器对第 i 类的基本概率分配函数 BPA[9-10] 为

$$m_i^k(A_i) = R_i^k p_i^k, \tag{2}$$

$$m_i^k(\Theta - A_i) = R_i^k(1 - p_i^k), \tag{3}$$

$$m_i^k(\Theta) = 1 - R_i^k. \tag{4}$$

其中: R_i^k 表示分类器 k 判定为第 i 类的准确率 (由训练数据得到, 且在半监督学习中是变化的); p_i^k 表示分类器 k 对该样本判定为 i 类的输出概率; A_i 表示“该样本属于第 i 类”, $(\Theta - A_i)$ 表示“该样本不属于第 i 类”, Θ 表示“该样本无法识别”. 因此, 对于每一个类, 均获得了 M 个包含相同焦元的 BPA. 由于 A_i 与 $(\Theta - A_j)$ $(i \neq j)$ 存在交集, 各 BPA 之间通常不会存在严重的冲突, 可采用 Dempster 规则组合.

为对待识别样本进行判决, 可进行如下 3 个步骤:

Step 1: 计算所有 R 个分类器对第 i 类的 BPA, 即

$$m_i(A_i) = \frac{\prod_{k=1}^{R}[m_i^k(A_i) + m_i^k(\Theta)] - \prod_{k=1}^{R} m_i^k(\Theta)}{1-K}, \tag{5}$$

$$m_i^k(\Theta - A_i) =$$

$$\frac{\prod_{k=1}^{R}[m_i^k(\Theta-A_i)+m_i^k(\Theta)]}{1-K}+\frac{-\prod_{k=1}^{R}m_i^k(\Theta)}{1-K},\quad(6)$$

$$m_i^k(\Theta)=\frac{\prod_{k=1}^{R}m_i^k(\Theta)}{1-K},\quad(7)$$

$$K=1-\prod_{k=1}^{R}[m_i^k(A_i)+m_i^k(\Theta)]+\prod_{k=1}^{R}m_i^k(\Theta)-\prod_{k=1}^{R}[m_i^k(\Theta-A_i)+m_i^k(\Theta)].\quad(8)$$

Step 2: 计算待识别样本对第 i 类总的 BPA, 即

$$m(A_i)=\Big\{m_i(A_i)\prod_{j=1,j\neq i}^{R}[m_j(\Theta-A_i)+m_j(\Theta)]+m_i(\Theta)\prod_{j=1,j\neq i}^{R}m_j(\Theta-A_i)\Big\}\Big/(1-K'),\quad(9)$$

$$m_i(\Theta-A_i)=\frac{m_i(\Theta-A_i)\prod_{j=1,j\neq i}^{R}m_j(\Theta)}{1-K'},\quad(10)$$

$$m_i^k(\Theta)=\frac{\prod_{j=1}^{R}m_j(\Theta)}{1-K'},\quad(11)$$

$$K'=-m_i(A_i)\prod_{j=1,j\neq i}^{R}[m_j(\Theta-A_i)+m_j(\Theta)]-\prod_{j=1}^{R}m_j(\Theta)-m_i(\Theta)\prod_{j=1,j\neq i}^{R}m_j(\Theta-A_i)-m_i(\Theta-A_i)\prod_{j=1,j\neq i}^{R}m_j(\Theta)+1.\quad(12)$$

Step 3: 显然, $\mathrm{Bel}(A_i)$ 等于 $m(A_i)$, 采用如下最大信任函数进行决策:

$$e=\begin{cases}j,\ \mathrm{Bel}(A_j)=\max\limits_{1\leqslant i\leqslant R}\mathrm{Bel}(A_i);\\ \text{rejected, otherwise.}\end{cases}\quad(13)$$

这样, 便可以选取判定出的样本并添加到训练数据集中, 然后采用 Zhou 等人[1]的 tri-train 算法进行半监督学习.

2.3 奇异标注点的剪辑

针对 Zhou 等人[1]提出的 tri-training 算法, 邓超等人[4]提出了 ADE-tri-training, 将基于最近邻规则的 Depuration 数据剪辑操作引入 tri-training 算法, 识别并移除每次迭代可能产生的误标记样例. 该方法实质上是采用 k 近邻方法, 观察 k 个近邻中是否有 k 个近邻的标记相同, 若不同, 则认为是“误标记”. 该方法具有运算简单、快速等优点. 然而, 当分析分布密度相差很大的数据时, 这种方法将会遇到困难. 基于以上考虑, 采用基于密度的局部异常因子检测算法 (LOF 算法) 进行再剪辑. LOF 算法考虑了对象的局域密度, 避免了基于距离的异常定义在数据集内部密度不一致的情况下出现的问题.

LOF 算法仅对新标记的样本进行操作, 下文描述的对象 p 就是新标记的样本. 下面介绍 LOF 算法的相关概念及其计算.

定义 1 对象 p 的 k-距离(k-distance). 对于任意的自然数 k, 定义 p 的 k-距离 (k-distance(p)) 为 p 与某个对象 o 之间的距离. 这里的 o 满足: 1) 至少存在 k 个对象 $o'\in D$, 使得 $d(p,o')\leqslant d(p,o)$; 2) 至多存在 $k-1$ 个对象 $o'\in D$, 使得 $d(p,o')<d(p,o)$.

定义 2 对象 p 的 k-距离邻域 (Nk-distance). 给定 p 的 k-距离 k-distance(p), p 的 MinPts-距离邻域包含所有与 p 的距离不超过 k-distance(p) 的对象

$$N_{k\text{-distance}}(p)=\{q|d(p,q)\leqslant k\text{-distance}(p)\}.\quad(14)$$

定义 3 对象 p 相对于对象 o 的可达距离. 给定自然数 k, 对象 p 相对于对象 o 的可达距离为

$$\mathrm{dist}_k(p,o)=\max\{k\text{-distance}(o),d(p,o)\},\quad(15)$$

这里 k-distance(o) 表示 p 在 o 的 MinPts 距离邻域内.

定义 4 对象 p 的局部可达密度为对象 p 与它的 MinPts 邻域的平均可达距离的倒数, 即

$$\mathrm{lrd}_{\mathrm{MinPts}}(p)=\frac{|N_{\mathrm{MinPts}}(p)|}{\sum\limits_{o\in N_{\mathrm{MinPts}}(p)}\mathrm{dis}\,t_{\mathrm{MinPts}}(p,o)}.\quad(16)$$

定义 5 对象 p 的局部异常因子 LOF 的计算公式为

$$\mathrm{LOF}_{\mathrm{MinPts}}(p)=\frac{\sum\limits_{o\in N_{\mathrm{MinPts}}(p)}\dfrac{\mathrm{lrd}_{\mathrm{MinPts}}(o)}{\mathrm{lrd}_{\mathrm{MinPts}}(p)}}{|N_{\mathrm{MinPts}}(p)|}.\quad(17)$$

对象 p 的局部异常因子 LOF 表示 p 的异常程度, 局部异常因子越大, 则认为它越可能异常; 反之, 则可能性小. 若 LOF 接近于 1, 则不被认为是局部异常, 而是处于簇的边缘或是簇的外面, 其对象的 LOF 相对较大.

LOF 算法实现步骤如下[11]:

Step 1: 确定参数 MinPts 的值, 对数据集中的每个对象计算其 MinPts-距离邻域, 并将该对象与邻域每个对象的距离存入数据库.

Step 2: 计算每个新标记对象的局部异常因子. 期间遍历数据集 2 次: 计算每个对象的局部可达密度; 计算每个新标记对象的局部异常因子.

Step 3: 对新标记对象 LOF 排序, 将最高的 n 个 LOF 对应的新标记对象剔除.

2.4 算法描述

在 tri-training 算法基础上, 先采用最小显著性差

异 (LSD) 方法挑选出显著性差异的成员分类器; 然后采用 D-S 证据理论标注; 再用 LOF 算法剔除误标记样本. 因此, 本文算法称为 LDL-tri-training 算法.

具体描述如下:

Step 1: 采用重采样技术生成若干个分类器.

Step 2: 采用最小显著性差异 (LSD) 方法挑选出 3 个分类器, 它们两两之间存在显著性差异.

Step 3: 用证据理论对未标记样本进行标注, 采用特定的 BPA 设定方式, 避免存在严重的冲突, 通过两阶段组合证据确定最大信任度进行标注.

Step 4: 采用 LOF 算法对新标记样本进行剪辑, 实现步骤如上节所述.

Step 5: 多次迭代, 检查是否满足 tri-training 算法的 3 个迭代标准条件: 1) 各个分类器的误差是否比上轮减小; 2) 被误标记的样本数是否也比上轮少; 3) 检查有标记数据集中样本数是否增加.

Step 6: 用这 3 个训练好的分类器进行多数投票预测.

3　实验结果与分析

选用 12 个 UCI 数据集进行实验, 见表 1. 对于每个数据集, 均取 20% 作为测试集, 剩余的 80% 作为训练集. 在训练集中分别按 20%, 40%, 60%和 80% 划分无标记样例集 U 和有标记样例集 L.

表 1　实验数据的构成信息

数据集	指标数	样本数	类别数
australian	14	690	2
buna	6	345	2
colic	22	368	2
diabetes	8	768	2
german	20	1 000	2
Ionosphere	34	351	2
hypothyroid	25	3 163	2
Kr-vs-kp	36	3 196	2
sick	29	3 772	2
Tic-tac-toe	9	958	2
vote	16	435	2
wdbc	30	569	2

为了更好地评估本文算法的性能, 与以下几种方法进行比较分析:

算法 1: 本节的 LDL-tri-training 算法;

算法 2: 文献 [1] 的 tri-training 算法;

算法 3: 文献 [4] 的 ADE-tri-training 算法;

算法 4: 最好的单分类器方法.

为了比较验证, 训练成员分类器的监督学习算法采用 weka 提供的 J48 决策树, 分类器选择中的 LSD 检验的置信度为 95%, 误标记样本剪辑的 LOF 中的 $k=3$. 对于每个数据集, 用上面所提到的 3 种方法构造大小为 3 分类器规模的融合, 并对不同无标记比例下每一种重复 5 次实验, 然后对每种方法计算其在 12 个数据集上的平均性能.

图 1 给出了不同无标记比例下各种方法的错误率平均值迭代变化过程. 从实验的情况看, 这 3 种方法的最大迭代次数均不超过 5 次, 且迭代 3 次后没有显著变化, 这一点也与文献 [4] 一致. 因此, 这 4 个子图均只显示了前 3 次的迭代情况.

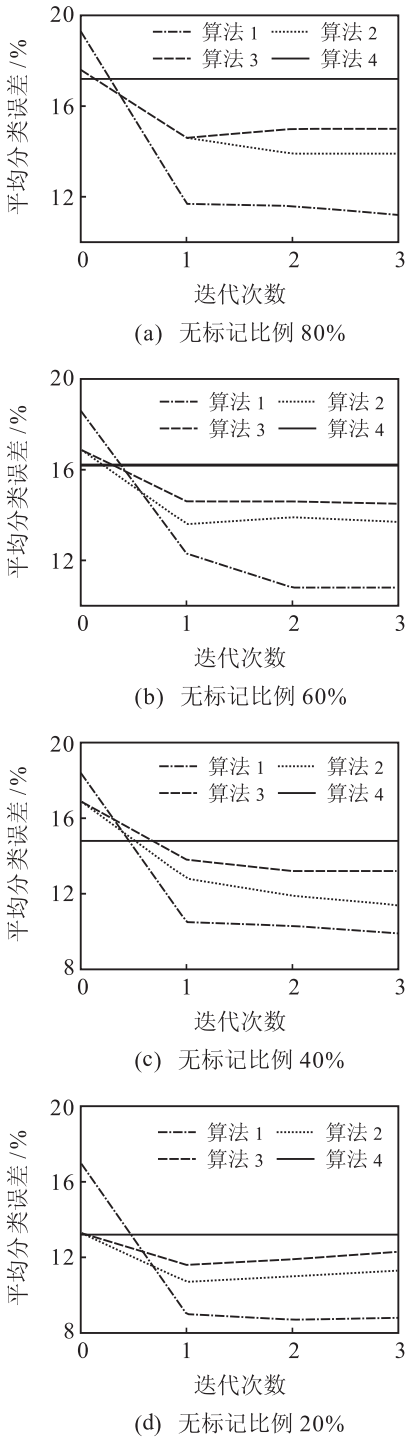

图 1　不同无标记比例下各方法错误率平均值迭代变化过程

从图 1 可以看出, 这 3 种算法与最好的单分类器方法相比, 在性能上均有很大提高, 说明均可以利用未标记样本提高泛化性能. 本文算法尽管在未迭代之

前性能最差(从而也说明成员分类器具有显著差异), 但迭代后性能提高非常明显, 从总的平均性能的提高情况看, 该算法效果最显著, 说明采用LSD检验的方法是有效的. 对比迭代过程可以看出, 本文算法的平均错误率基本上是持续下降, 而其他2种方法会出现波动, 甚至轻微上扬, 说明采用LOF算法剔除误标记样本是有效的.

图2表示有标记只占极小比例时平均分类误差变化过程. 可以看出, 随着有标记样本的减少, 3种算法平均误差率均急剧增加. 但是, 本文算法增加的幅度远远小于其他2种算法, 可能的原因是采用统计证据对未标记样本标注, 而不是采用其他2种方法的多数投票法.

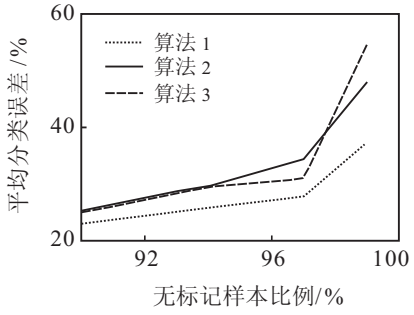

图 2 有标记占极小比例时平均分类误差变化过程

综上可以看出, 本文算法具有更好的稳定性, 在迭代过程中受误标记噪声影响最小, 能够显著提高泛化性能.

4 结 论

本文在前人众多研究成果的基础上, 提出了一种新的LDL-tri-training算法. 该算法通过最小显著性差异(LSD)假设检验方法使得3个成员分类器两两之间具有显著性差异; 然后采用D-S证据理论标注提高标注的稳定性; 最后利用LOF算法剔除误标记的噪声样本. 实验表明, LDL-tri-training算法在无标记样本所占比例为20% ∼ 80%时均具有较高的分类精度和稳定性, 即使在有标记样本极少的情况下, 也具有不错的表现.

参考文献(References)

[1] Zhou Zhihua, Li Ming. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[2] 王珏, 周志华, 周傲英. 机器学习及其应用[M]. 北京: 清华大学出版社, 2007: 259-275.
(Wang J, Zhou Z H, Zhou A Y. Machine learning and application[M]. Beijing: Tsinghua University Press, 2007: 259-275.)

[3] Li Ming, Zhou Zhihua. SETRED: Self-training with editing[C]. Proc of the 9th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Hanoi: Springer, 2005: 611-621.

[4] 邓超, 郭茂祖. 基于自适应数据剪辑策略的Tri-training算法[J]. 计算机学报, 2007, 30(8): 1213-1226.
(Deng C, Guo M Z. ADE-Tri-training: Tri-training with adaptive data editing[J]. Chinese J of Computers, 2007, 30(8): 1213-1226.)

[5] Wang Wei, Zhou Zhihua. Analyzing co-training style algorithms[C]. Proc of the 18th European Conf on Machine Learning. Warsaw, 2007: 454-465.

[6] Nadeau C, Bengio Y. Inference for the generalization error[J]. Machine Learning, 2003, 52(3): 239-281.

[7] David R Anderson, Dennis J Sweeney. Statistics for business and economics[M]. Beijing: China Machine Press, 2010: 327-330.)

[8] Dymitr Ruta, Bogdan Gabrys. An overview of classifier fusion methods[J]. Computing and Information Systems, 2000, 7(1): 1-10.

[9] 蔡自兴, 徐光佑. 人工智能及其应用[M]. 北京: 清华大学出版社, 2004: 96-100.
(Cai Z X, Xu G Y. Artificial intelligence and application[M]. Beijing: Tsinghua University Press, 2004: 96-100.)

[10] 匡小新, 王先甲. 水文特性参数的证据统计推断方法[J]. 武汉大学学报, 2003, 36(1): 32-36.
(Kuang X X, Wang X J. An evidential statistical inference method of hydrologic characteristic parameters[J]. J of Wuhan University, 2003, 36(1): 32-36.)

[11] 杨风召, 朱扬勇, 施伯乐. IncLOF: 动态环境下局部异常的增量挖掘算法[J]. 计算机研究与发展, 2004, 41(3): 477-484.
(Yang F Z, Zhu Y Y, Shi B L. IncLOF: An incremental algorithm for mining local outliers in dynamic environment[J]. J of Computer Research and Development, 2004, 41(3): 477-484.)

[12] Xiao-liang Tang, Min Han. Ternary reversible extreme learning machines: The incremental tri-training method for semi-supervised classification[J]. Knowledge and Information Systems, 2010, 23(3): 345-372.