一种基于证据理论的多类半监督分类算法

盛 凯[1],刘 忠[1],周德超[1],魏启航[2],冯成旭[1]
(1. 海军工程大学兵器工程学院,湖北武汉 430033;2. 66029 部队,内蒙古锡林郭勒盟 011216)

摘 要: 为了提高多类半监督分类的性能,提出了一种基于证据理论的多类协同森林算法(DSM-Co-Forest).首先,通过"多对多"模式将有标记的多类数据随机拆分为多个二类数据集,并以此训练二类基分类器;然后,利用多个基分类器同时对未标记样本进行预测,并利用证据组合算法挑选出可信度较高的未标记样本;最后,将高可信度的未标记样本加入到原训练样本中,以迭代更新其他的基分类器,从而提高分类器的整体性能.通过在一些公共数据集上进行实验,并与其他半监督分类算法进行对比,验证了所提算法的可行性和有效性.

关键词: 半监督学习; 多类分类; 证据理论; 协同森林
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2018)11-2642-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.11.011

A Multi-class Semi-Supervised Classification Algorithm Based on Evidence Theory

SHENG Kai[1], LIU Zhong[1], ZHOU De-chao[1], WEI Qi-hang[2], FENG Cheng-xu[1]
(1. *College of Weapons Engineering, Naval University of Engineering, Wuhan, Hubei* 430033, *China*;
2. *PLA* 66029 *Troop, Xilinguole, Inner-Mongolia* 011216, *China*)

Abstract: In order to improve the performance of multi-class semi-supervised classification, a new multi-class Co-Forest algorithm named DSM-Co-Forest is proposed on the basis of D-S evidence theory. First, through MVM mode, the multi-labeled data set is randomly split into multiple binary-class data set to train the base classifiers; then, these base classifiers are used to pick out the high reliability samples from the unlabeled data set by using the evidence combination algorithm; finally, adds these selected samples to the original training set to iteratively update the base classifiers so as to improve the overall performance of the multi-class classifier. Through comparing with other semi-supervised classification algorithms on several public data sets, the feasibility and validity of the proposed algorithm are verified.

Key words: semi-supervised learning; multi-class classification; evidence theory; co-forest.

1 引言

半监督分类是当前机器学习领域的一个重要研究方向[1].区别于传统的有监督分类算法,它试图通过利用大量廉价的未标记数据来辅助少量的有标记数据进行学习,从而有效提高分类器的泛化性能.按学习方式不同,常见的半监督分类算法可大致分为四类:基于生成模型的算法、基于支持向量机的算法、基于图的算法以及基于分歧的算法[2].其中,基于分歧的算法通过训练多个"显著差异"的基分类器来对未标记样本进行标记,然后通过相互迭代的方式完成各基分类器的更新.由于其受到模型假设影响少、学习方法简单、理论基础坚实以及适用范围广泛等优点[3],该类算法近年来受到较大的关注和发展.

基于分歧的半监督分类算法起始于 1998 年 Blum 等人提出的 Co-Training 算法[4].该算法假设数据拥有两个充分且条件独立的视图,首先基于有标记样本在每个数据视图上分别训练一个分类器,然后从未标记样本中标注各分类器认为的高可信度样本,并将其用于另一个分类器的更新.但是,现实中的数据大多难以满足该算法的假设条件.为了解决该问题,Zhou 等人先后提出了基于三个和多个基分类器进行协同训练的 Tri-Training 算法[5]和 Co-Forest 算法[6],分别通过在单个视图上训练有差异的三个或多个基分类器,再采用

收稿日期:2017-12-03;修回日期:2018-04-22;责任编辑:蓝红杰
基金项目:湖北省自然科学基金(No. 2017CFB377)

多数投票法标注未标记样本,并将其加入到有标记样本集中对其他的基分类器进行更新;此后,基于分歧的半监督分类算法得到进一步发展,例如,Deng 等人提出的 ADE-Co-Forest 算法[7]、于重重等人提出的 DSCC 算法[8]、陈思等人提出的 Co-SemiBoost 算法[9]、张磊等人提出的结合聚类特征的半监督协同分类算法[10]、邹保平等人提出的改进 Co-Forest 算法[11]以及 Jia 等人提出的 M-Training 算法[12]等,这些算法在医学诊断、故障检测及图像识别等多个研究领域得到广泛应用.

然而,上述研究主要集中在二分类问题上,并不能够很好应对现实中更广泛存在的多分类问题. 本文受文献[13~16]的启发,将 D-S 证据理论与 Co-Forest 算法相结合,提出了一种基于证据理论的多类半监督分类算法(DSM-Co-Forest). 首先,按照随机组合的方式将多类数据划分为多个两类数据,用以训练初始基分类器;其次,采用 D-S 证据理论将部分基分类器对未标记样本的分类结果进行融合,并将高可信度样本加入到其余的基分类器的训练集中;然后,令各基分类器之间相互迭代更新,直至达到终止条件. 实验结果表明,本文提出的算法有效提高了多类半监督问题的分类正确率,并具有较高的稳定性.

2　相关工作

2.1　Co-Forest 算法

Co-Forest 算法将 Tri-Training 算法与随机森林(Random Forest)分类算法相结合,利用集成学习的优势,将标注未标记样本中的高可信度样本的基分类器的个数由 Tri-Training 算法中的 2 个扩展为 N-1 个(N 为基分类器个数),并考虑了样本的可信度权重. 算法的核心步骤可概述如下:

Step 1. 采用随机子空间和抽样方法,利用有标记数据集 L 习得 N 个基分类器,由此构成初始分类器 $H^*=\{h_1,h_2,\cdots,h_N\}$;

Step 2. 令 H_i 为除了基分类器 h_i 之外的其它基分类器集合($H_i=H^*-h_i$),利用袋外数据(out of bag)估计 H_i 在第 t 轮训练后的分类误差 $err_{i,t}$. 若 $err_{i,t}<err_{i,t-1}$,则根据 H_i 挑选未标记数据集 U 中的高可信度样本,并将其加入到 h_i 的原训练集中,训练更新基分类器 h_i;

Step 3. 重复 Step 2,直到所有基分类器都不再满足更新条件.

Co-Forest 算法通过利用随机抽样和随机子空间等方法尽可能使各基分类器之间具有差异性(独立性),并通过多个基分类器进行投票来提高伪标记样本的准确度,从而改善分类器性能. 该算法能够较好的应对二分类问题,但是在处理多分类问题时将会出现以下问题:(1)分类器所集成的弱(基)分类器的分类正确率普遍由二分类时的 $1/2+r$ 降低为 $1/k+r$(k 为类别数,$r>0$),因此需要更多的基分类器进行投票,且在未标记样本集中很难获得具有较高可信度的样本;(2)在多分类器中,存在训练样本分到正确类的概率小于错分到某一错误类的概率的现象[17],这可能使部分被错误标注的未标记样本参与到分类器的更新训练中,从而降低分类器的性能.

2.2　D-S 证据理论

D-S 证据理论在信息融合、模式识别以及决策分析等领域已得到广泛应用[18,19],其基本概念如下:

(1)基本概率分配、信任函数和似然函数

假设全域 $X=\{x_1,x_2,\cdots,x_n\}$ 为系统中所有可能发生的状态,且 X 的各个元素互斥,则 X 称为辨识框架. 令 2^X 表示 X 的所有子集组成的幂集,则任何一个命题 A 可以表达为 2^X 的一个子集.

设 X 为辨识框,函数 $m:2^X\to[0,1]$ 称为基本概率分配函数(basic probability assignment,BPA),表示对命题 A 的精确信任度. 对于空集 φ,$m(\varphi)=0$;对于 $A\subseteq 2^X$,$\sum m(A)=1$. 若 $m(A)>0$,则称 A 为该函数的一个焦元.

函数 $Bel:2^X\to[0,1]$ 称为信任函数,表示对 A 的全部信任. 其中,$Bel(A)=\sum_{B\subseteq A}m(B)$.

函数 $Pl:2^X\to[0,1]$ 称为似然函数,表示对 A 的潜在信任. 其中,$Pl(A)=\sum_{A\cap B\neq\varnothing}m(B)$.

(2)Dempster 组合规则

设 m_1 和 m_2 分别是 X 上的两个独立证据的概率分配函数,则它们的 *Dempster* 组合结果也是一个概率分配函数,其定义为:

$$m(A)=\begin{cases}0, & A=\varnothing\\ c^{-1}\sum_{x\cap y=A}m_1(x)m_2(y), & A\neq\varnothing\end{cases}\tag{1}$$

其中,$c=1-\sum_{x\cap y=\varphi}m_1(x)m_2(y)$.

类似的,如果有 n 个相互独立的证据,其 Dempster 组合结果为:

$$m(A)=\begin{cases}0, & A=\varnothing\\ c^{-1}\sum_{\cap A_i=A}\prod_i m_i(A_i), & A\neq\varnothing\end{cases}\tag{2}$$

其中,$c=1-\sum_{\cap A_i=\varnothing}\prod_i m_i(A_i)=\sum_{\cap A_i\neq\varnothing}\prod_i m_i(A_i)$.

3　结合证据理论的半监督分类算法

3.1　基于证据理论的多分类算法框架

不同于常见的“一对一”(OVO)或“多对多”(OVR)多分类模式,本文采用“多对多”(MVM)数据拆分策略和基于证据理论的信息融合方法进行多类分类,其基本流程如图 1 所示.

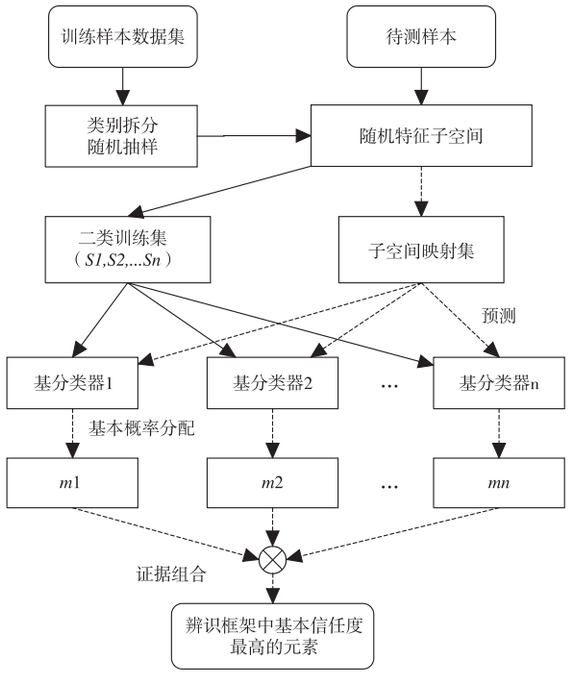

图1 基于证据理论的多类分类流程

图1中,实线表示分类器的训练过程,虚线表示新样本的预测过程. 在分类器的训练过程中,首先需要对多类数据集进行 n 次拆分,每次随机抽取若干类别作为正类,其他类作为负类,从而共形成 n 个二类数据集;然后,采用能够返回预测概率的二类分类算法训练得到 n 个基分类器. 在预测过程中,首先根据每个基分类器对未知样本的预测概率构造 n 个基本概率分配(BPA)函数$\{m_1, m_2 \cdots, m_n\}$;然后,根据 Dempster 证据组合规则,将获得的信任度最高的类别做为最终的预测结果.

下面从理论上证明上述基于证据理论的多分类算法框架的有效性.

假设训练样本中类别总数为 k,根据 *D-S* 证据理论,可定义识别框架为 $Y=\{y_1, y_2 \cdots, y_k\}$. 令类别在第 i 次拆分后得到的两个子类分别是 A_{i1} 和 A_{i2},则有 $A_{i1} \subset Y$, $A_{i2} \subset Y$,且 $A_{i1}=Y-A_{i2}$. 因此,基分类器 i 在对样本 x 进行预测时,可得属于两个类别的概率分别为 p_i 和 $1-p_i$. 假设该基分类器的分类正确率为 r_i,则可定义第 i 个基分类器对样本 x 所属类别的 BPA 函数为:

$$\begin{cases} m_i(A_{i1})=r_i p_i, \\ m_i(A_{i2})=r_i(1-p_i), \\ m_i(Y)=1-r_i. \end{cases} \text{或} \begin{cases} m_i(A_{i1})=p_i, \\ m_i(A_{i2})=1-p_i. \end{cases} \tag{3}$$

若根据上述方法将训练样本随机拆分 n 次,将可通过训练得到 n 个基分类器;根据这些基分类器对未知样本的预测结果,利用式(3)构造出 n 个 BPA 函数. 根据 D-S 证据理论可得出以下推论:

推论 1 若构证据时考虑基分类器的正确率,即利用式(3)的左式构造 BPA 函数,如果所有证据相互独立且基分类器的平均正确率 r 满足 $0.5<r\leq 1$,则当 n 足够大时,所获得的未知样本的全部证据经 Dempster 组合后,其真实类别所在焦元的 BPA 一定存在,且在所有单元素焦元中其期望值最大.

证明 经类别拆分后,识别框架中的任意单独元素 $y_a \in Y$ 存在且仅存在于其中的一个子类中,即 $y_a \in A_{i1}$ 且 $y_a \notin A_{i2}$,或者 $y_a \notin A_{i1}$ 且 $y_a \in A_{i2}$. 若考虑分类器正确率的影响,并假设在每次拆分后 $y_a \in A_{ix}$,则当类别随机拆分的次数 n 足够大时,一定存在 $\cap A_{ix}=y_a$. 因此,所有 BPA 函数经 Dempster 组合后,任意单独元素 y_a 的焦元一定存在,且 $m(y_a)>0$.

不妨假设未知样本 x 的真实类别为 y_c,则 y_c 在子类 A_{i1} 的概率为 r(此时该基分类器预测正确),在子类 A_{i2} 的概率为 $1-r$(此时该基分类器预测错误). 根据式(2)可得,证据组合后 y_c 的 BPA 期望为:

$$E[m(y_c)]=c^{-1}\cdot(2^n-1)\cdot(rp)^{\frac{1}{2}nr} \cdot(r-rp)^{\frac{1}{2}n(1-r)}(1-r)^{\frac{1}{2}n} \tag{4}$$

对于其他 $y_a \neq y_c$,y_a 被分到任一子类的概率相同,因此,有:

$$E[m(y_a)]=c^{-1}\cdot(2^n-1)(rp)^{\frac{1}{4}n}(r-rp)^{\frac{1}{4}n}(1-r)^{\frac{1}{2}n} \tag{5}$$

由于 $rp>r-rp$ 且 $r>0.5$,因此,由式(4)、(5)可得 $E[m(y_c)]>E[m(y_a)]$,即经证据组合后,真实类别所在焦元的 BPA 期望值大于其他单元素焦元的 BPA 期望值. 得证.

推论 2 若构造证据时不考虑基分类器的正确率,即利用式(3)中的右式构造 BPA 函数,如果所有证据相互独立,且基分类器的平均正确率 r 满足 $0.5<r\leq 1$,则当 n 足够大时,所有证据经 Dempster 组合后,未知样本的真实类别所在焦元的 BPA 期望值将趋近于 1.

证明 若不考虑基分类器的平均正确率,则对于任意 $y_a \in Y$,每个 BPA 函数 m_i 中有且只有一个焦元包含 y_a. 因此,当 n 足够大时,BPA 函数集中有且仅有 k 个单独的类别构成的焦元满足 $\cap A_i \neq \varnothing$. 因此,由根据式(2)可得,经 Dempster 组合后 y_a 的 BPA 为:

$$m(y_a)=\frac{\prod_i m_i(y_a)}{\sum_k \prod_i m_i(y_a)} \tag{6}$$

不妨假设样本 x 的真实类别为 y_c,当基分类器分类正确时,给出 y_c 所在子类别的概率为$(1/2+\varepsilon)$,否则为$(1/2-\varepsilon)$. 假设每个基分类器的正确率为 r,则平均有 $n\cdot r$个基分类器分类正确. 因此,有:

$$E[\prod_i m_i(y_c)] = (1/2+\varepsilon)^{nr} \cdot (1/2-\varepsilon)^{n-nr} \quad (7)$$

对于其他 $y_a \neq y_c$,y_a 与 y_c 被划分为同一子类的概率为 $p_{a,c} = \left(\frac{k-1}{2k}\right)$,因此,有:

$$\begin{cases} E[\prod_i m_i(y_a)] = (1/2+\varepsilon)^P \cdot (1/2-\varepsilon)^Q \\ P = p_{a,c} \cdot nr + (1-p_{a,c}) \cdot (n-nr) \\ Q = p_{a,c} \cdot (n-nr) + (1-p_{a,c}) \cdot nr \end{cases} \quad (8)$$

求 $m(y_c)$ 与 $m(y_a)$ 的比值,可得:

$$E\left[\frac{m(y_c)}{m(y_a)}\right] = \left(\frac{1/2+\varepsilon}{1/2-\varepsilon}\right)^{\frac{(2r-1)\cdot(1+k)}{2k} \cdot n} \quad (9)$$

由于 $0 < \varepsilon \leq 0.5$,$0.5 < r \leq 1$,当 n 足够大时,由式(9)易得 $E[m(y_c)] > E[m(y_a)]$,即经证据组合后,元素 y_c 的 BPA 值最大.

将式(7)、(8)代入式(6),可得:

$$E[m(y_c)] = \frac{E[\prod_i m_i(y_c)]}{(k-1)E[\prod_i m_i(y_a)] + E[\prod_i m_i(y_c)]} \quad (10)$$

化简,得:

$$\begin{cases} E[m(y_c)] = \left[1 + (k-1) \cdot \left(\frac{1/2+\varepsilon}{1/2-\varepsilon}\right)\exp(Z)\right]^{-1} \\ Z = \frac{(2r-1)\cdot(1+k)}{2k} \cdot n \end{cases} \quad (11)$$

由式(11)可得,随着基分类器数量 n 的增加,$m(y_c)$ 的期望将逐渐逼近于 1. 得证.

理论上,基分类器的个数越多,分类的正确率越高,但随着 n 的不断增大,其逼近速度在迅速降低. 同时,由于过多的基分类器会导致计算量过大,且不利于保证基分类器之间的独立性,因此,一般取 n 为 $[k,3k]$ 之间的一个整数即可.

3.2　DSM-Co-Forest 算法

针对多类半监督分类问题,本节将 Co-Forest 算法思想与 3.1 节提出的基于证据理论的多分类算法框架相结合,提出一种基于证据理论的多类协同森林算法(D-S evidence theory based Multi-class Co-Forest 算法,简称 DSM-Co-Forest). 该算法首先利用有标记样本训练多个二分类基分类器;然后基于 D-S 证据理论标注高可信度的未标记样本;再利用伪标记样本更新基分类器. 算法的具体描述如算法 1 所示.

算法 1　DSM-Co-Forest

输入:有标记训练集 L,未标记训练集 U,基分类器个数 n,可信度阈值 t
输出:半监督分类器 H
步骤:

1. 利用分层抽样法将 L 划分为 *L_train* 和 *L_estimate* 两部分;
2. 采用 *Bootstrap* 抽样法和随机子空间法,从 *L_train* 中获得 n 个有标记子数据集 $\{L_1, L_2, \cdots, L_n\}$;
3. 对每个有标记子训练集随机拆分成二类数据,训练 n 个二类基分类器 $H = \{h_1, h_2, \cdots, h_n\}$;
4. 初始化基分类器的本轮估计误差变量 err_i、上轮估计误差变量 err_prime_i;
5. 对每个基分类器 h_i,执行:
5.1 利用 H_i 和 *L_estimate* 估计 err_i. 其中,$H_i = H - h_i$;
5.2 若 $err_i < err\_prime_i$,则从 U 中随机抽取子集 U_i',并计算 U_i' 中每个样本 x_u 的可信度 t_u. 若 $t_u > t$,则 $L_i = L_i + x_u$,并更新 h_i;
6. 剔除 U 中所有被挑选出的高可信度样本;
7. 重复执行步骤 5、6,直到没有基分类器满足更新条件;
8. 基于训练后的 n 个基分类器获得证据,并通过证据组合来预测新样本的类别.
结束.

算法首先利用分层抽样法在全部有标记训练集中抽取部分样本 *L_estimate* 作为独立的评估样本(第 1 步),不参与任何基分类器的训练,从而保证了对各基分类器性能的有效评估;在初始基分类器的训练中(第 2、3 步),算法采用了随机抽样、随机子空间以及随机拆分等方法构建有标记子训练集,以最大程度的保证各基分类器之间的相对独立性. 基分类器需选择能够获取各类预测概率值的分类模型;在协同训练过程中(第 5~7 步),令 H_i 为除 h_i 之外的其他所有基分类器($H_i = H - h_i$),利用独立评估样本 *L_estimate* 对 H_i 的错误率 err_i 进行估计. 如果 $err_i < err\_prime_i$,说明当前轮 H_i 性能优于上一轮,因此有可能挑选出更多可信的未标记样本用于 h_i 的更新,其中,样本的可信度可通过 Pignistic 概率转换[20]方法计算,如式(12)、(13)所示.

$$P_m(y_i) = \sum_{x_i \in B} \frac{m(B)}{|B|}, \ \forall B \subseteq X \quad (12)$$

$$t_u = \max_{1 \leq i \leq k} \{P_m(y_i)\} \quad (13)$$

在 Co-Forest 及改进算法中,为保证各轮次和各基分类器挑选的新样本之间具有差异性,高可信度样本只从 U 的一个子集中进行评估,子集样本个数的约束如式(14)所示. 其中,W_prime_i 为上一轮 H_i 所标注样本的可信度累计之和. 但是该方法往往导致挑选的样本个数较少,没有对未标记样本进行充分利用. 本文算法放松了对待选样本的个数限制条件,直接从 U 中抽取一定比例的样本参与可信度评估,如 $|U_i'| = |U|/n$. 同时在本轮次训练结束后,将所有被标注的样本从 U 中剔除(第 6 步),以保证参与分类器更新的未标记样本不会在下一轮训练中再被选中.

$$|U'| \leq \frac{err\_prime_i \cdot W\_prime_i}{err_i} \quad (14)$$

算法中还有一个关键参数——可信度阈值 t. 该值过高,可能导致被选中的样本太少,限制了分类器的进一步

提升;该值过低,则可能导致部分错误标记的样本参与基分类器的更新,反而降低分类器的性能. 在算法的实际应用中,可以根据分类器在独立评估样本上的预测准确度进行调整. 这里仅给出一个合理的估计:阈值参数 t 等于证据组合后正确类别的基本信任度的期望,即 $t=\max_{1\le i\le k}E[m(y_i)]$. 当取 $n=2k$ 时,由式(11)可得:

$$t=\frac{1}{1+(k-1)\cdot\left(\frac{1/2+\varepsilon}{1/2-\varepsilon}\right)\exp((2r-1)\cdot(1+k))} \tag{15}$$

其中,r 和 ε 可由 *L_train* 训练得到的初始基分类器在 *L_estimate* 进行估计获得,r 为所有基分类器正确率的平均值,$(1/2+\varepsilon)$为所有预测正确的样本其预测概率的平均值.

4 实验

4.1 数据集

本文选取 8 个 UCI 标准数据集[21]进行测试,分别是 German、Glass、Wine、Wifi-Localization、Vehicle、Letter、Abalone 和 Segment. 为了便于分析,实验前首先对样本的属性进行归一化,并进行以下处理:将 Glass 数据集的类别分为浮法处理的窗玻璃(float processed window glass)、未经浮法处理的窗玻璃(non-float processed window glass)以及非窗玻璃(non-window glass)等 3 类;从 Letter 数据集中提取“A”、“B”、“C”、“D”等 4 类样本数据;从 Abalone 数据集中提取年轮数小于4、等于6、等于9、等于 12、等于 15 以及大于 18 等 6 类样本数据. 数据集的总体描述如表 1 所示.

表 1 数据集总体特性描述

数据集	属性数	类别数	样本数
German	23	2	1000
Glass	9	3	214
Wine	13	3	178
Wifi-Localization	7	4	2000
Vehicle	18	4	846
Letter	16	4	3096
Abalone	7	6	1486
Segment	19	7	2310

4.2 实验对比

本文所提算法通过 Python 进行实现,并与 Tri-Training[5]、Co-Forest[6]及 DS-Biased-SVM[15]算法进行比较,同时选取有监督分类算法 Random Forest 作为基准. 为了满足多类分类的需求,本文选择 scikit-learn 机器学习包中基于 OVR 模式扩展的 SVM 多分类器作为 Tri-Training、Co-Forest 以及 DS-Biased-SVM 算法的基分类器,并采用 Platt 概率估计模型[22]获取各类的概率值. 同时,设置 Tri-Training、Co-Forest 算法的可信度阈值为 0.9;设置 Co-Forest 算法及所提算法的基分类器个数为样本类别数的 2 倍;不同算法中基分类器的参数根据训练集进行调节,使得各算法的分类效果达到最优.

实验中,选取每个数据集的 20% 作为测试集,其余的 80% 作为训练集;然后再从训练集中采用分层随机抽样的方式分别从中选取 10%、30% 和 50% 作为有标记数据集,其余的作为无标记数据集. 实验以分类器在测试集上的总体分类正确率作为评价标准,并通过 100 次重复实验取平均值,其结果如表 2 所示.

表 2 不同半监督分类算法的分类正确率对比

数据集	标记样本比例	Random Forest	Tri-Training	Co-Forest	DS-Biased-SVM	Proposed
German	10%	0.6835	0.6955	**0.7124**	0.7000	0.7025
	30%	0.7015	0.7130	**0.7255**	0.7015	0.7220
	50%	0.7221	0.7240	0.7398	0.7047	**0.7431**
Glass	10%	0.7942	0.8104	0.8190	0.8048	**0.8254**
	30%	0.8219	0.8235	0.8253	0.8109	**0.8452**
	50%	0.8507	0.8719	0.8785	0.8566	**0.8928**
Wine	10%	0.9158	0.9764	**0.9972**	0.9586	0.9861
	30%	0.9358	**0.9947**	0.9722	0.9911	0.9805
	50%	0.9575	0.9997	**1.0000**	0.9972	0.9888
Wifi-Localization	10%	0.9572	0.9642	0.9700	0.9678	**0.9728**
	30%	0.9705	0.9733	0.9746	0.9720	**0.9750**
	50%	0.9775	0.9790	**0.9813**	0.9797	0.9800
Vehicle	10%	0.6131	0.6681	0.6725	0.6642	**0.6889**
	30%	0.7015	0.7486	0.7573	0.7520	**0.7591**
	50%	0.7246	0.8041	0.8156	0.8035	**0.8327**
Letter	10%	0.9330	0.9512	0.9547	0.9544	**0.9583**
	30%	0.9608	0.9793	0.9812	0.9748	**0.9822**
	50%	0.9813	0.9875	0.9877	0.9835	**0.9903**

续表

数据集	标记样本比例	Random Forest	Tri-Training	Co-Forest	DS-Biased-SVM	Proposed
Abalone	10%	0.5274	0.6045	0.6164	0.5856	**0.6187**
	30%	0.6036	0.6474	0.6662	0.6413	**0.6789**
	50%	0.6160	0.6366	0.6502	0.6324	**0.6582**
Segment	10%	0.9217	0.9333	0.9283	0.9274	**0.9372**
	30%	0.9517	0.9568	0.9553	0.9456	**0.9575**
	50%	0.9590	0.9560	0.9526	0.9552	**0.9653**
获胜次数		0	1	5	0	**18**

从表 2 中可以看出,相对于只用有标记样本训练的 Random Forest 分类器,半监督算法的分类效果均有所提高,且随着标记样本比例的增加而升高. 当样本的类别较少时,本文所提算法与 Co-Forest 算法不相伯仲,但是当样本类别较多($k>3$)时,本文算法在总体分类正确率上更具有优势. 这是因为相对于传统的基于 OVO 或 OVR 等拆分的多分类算法,本文所提算法的每个基分类器仅对多类样本做一次 MVM 拆分,避免了多次拆分导致的误差累积问题;在对未标记样本的类别进行预测时,基于证据理论的信息融合算法不但考虑了预测正确的可能性,同时也考虑了预测错误及不确定的可能性,在对各个基分类器预测信息的利用上比(加权)投票法更加充分,因此进行伪标记的未标记样本可信度更高.

4.3　算法参数对分类性能的影响

本文在第 3 节中,从理论上分析了基分类器个数 n 和可信度阈值 t 对半监督学习的影响,本小节通过人工数据集进行实验验证.

(1)基分类器个数 n 的影响

通过 scikit-learn 机器学习包提供的分类数据生成器构造特征数为 10,类别数(k)分别为 3、4、5、6,样本总数为 1000 的四个数据集,分别选取其中的 80% 作为训练样本集,20% 作为测试样本集;然后选择训练样本集中的 30% 作为有标记样本,70% 作为未标记样本训练分类器,然后在测试样本集上进行测试. 通过 20 次实验求取平均值,实验结果如图 2 和图 3 所示.

图 2 和图 3 分别为基分类器个数 n 对分类正确率和学习耗时的影响. 由图 2 可以看出,随着基分类器个数的增加,分类正确率总体上在不断地提高,但是提高幅度逐渐降低,这与本文第 3.1 节的理论分析结果相一致. 由图 3 可以看出,当基分类器的个数增多时,学习耗时迅速上升. 这不但是因为在每一轮训练中需要更新的基分类器个数增多了,而且由于更多的基分类器可能会发现更多的高可信度样本,从而导致训练的轮次也相应增加. 因此,在实际的半监督学习中,基分类器的个数需要在分类正确率和学习耗时之间折衷考虑.

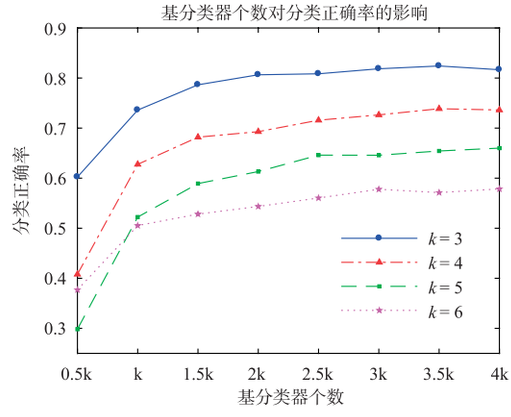

图 2　基分类器个数对分类精度的影响

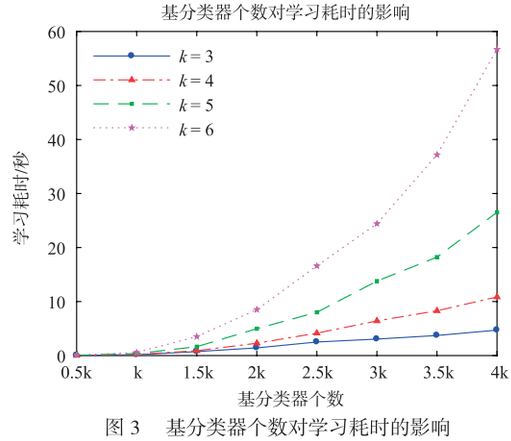

图 3　基分类器个数对学习耗时的影响

(2)可信度阈值 t 的影响

根据上述方法构造特征数为 10,类别数(k)为 4,样本总数为 1000 的数据集,选取其中的 80% 作为训练样本集,20% 作为测试样本集;然后选择训练样本集中的 30% 作为有标记样本,70% 作为未标记样本训练分类器,然后在测试样本集上进行测试. 基分类器个数固定为 $n=10$,可信度阈值的取值范围设定为 0.6 ~ 1.0,通过 20 次实验求取平均值,实验结果如图 4 所示.

图 4 中,蓝色实线表示分类器的预测正确率,红色虚线表示未标记样本中未参与训练的比例. 由图中可以看出,随着 t 的升高,未参与训练的未标记样本比例也随之升高,当可信度阈值 $t=1.0$ 时,未参与训练的未

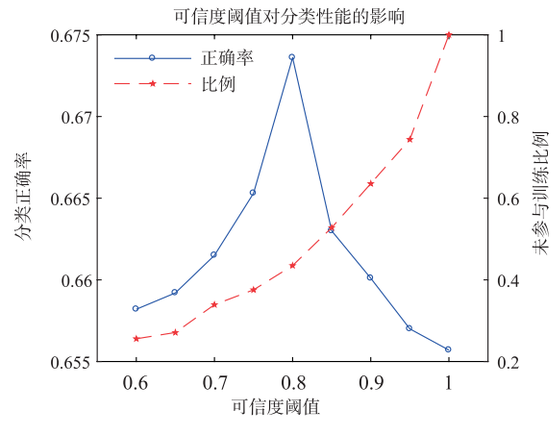

图4 可信度阈值对分类性能的影响

标记样本比例也为1.0,即此时没有满足条件的未标记样本能够参与到分类器的训练中. 同时,分类器的分类正确率随着 t 的增加先升高后降低,当 $t=0.8$ 时达到最高值. 这是由于当 t 较高时,只有少数未标记样本被选中,限制了分类器性能的进一步提升;而当 t 较低时,存在部分被错误标注的未标记样本参与到了分类器的训练中,从而导致分类器性能的降低;此外,当 $t=1.0$ 时,由于没有未标记样本参与训练,此时分类器仅由有标记样本训练所得,其性能明显低于本文所提的半监督分类算法.

5 结论

针对现实数据中存在大量未标记样本未被充分利用的问题,本文将信息融合领域的D-S证据理论与基于分歧的半监督分类算法Co-Forest相结合,提出了一种新的多类半监督分类算法DSM-Co-Forest. 相较其他基于分歧的半监督分类算法,本文所提算法具有以下优势:(1)基分类器直接采用更加成熟的二分类器模型,基分类器的预测正确率更高;(2)在未标记样本的标记过程中,采用无放回的等比例抽样法,更好的保证了基分类器之间差异性;(3)采用证据组合法将多个基分类器的预测结果进行融合,对信息的利用更加充分. 通过在UCI数据集上和其他半监督分类方法的对比实验,表明了本文所提算法在解决多类半监督分类问题上的有效性.

本文所提算法也有一定的不足. 一是样本的类别数较多时,需要训练的基分类器个数随之增加,同时训练轮次也相应增多,这将导致分类器的训练耗时显著增长. 如何在保证分类准确率的同时有效减少训练时间是一个难点;二是对于可信度阈值的取值,本文虽然进行了一定的理论分析,但是在实际中该参数的取值仍需要进行人工调试. 在以后的工作中可以探索根据数据样本的分布来自动确定参数值;三是本文算法没有考虑多类样本的数据不平衡问题. 目前,已有多种针对不平衡数据的分类算法,如何将其结合在多类半监督分类问题中,还有待于进一步研究.

参考文献

[1] 刘建伟,刘媛,罗雄麟. 半监督学习方法[J]. 计算机学报,2015,38(8):1592-1617.
LIU Jian-wei, LIU Yuan, LUO Xiong-lin. Semi-supervised learning methods[J]. Chinese Journal of Computers, 2015, 38(8):1592-1617. (in Chinese)

[2] 周志华. 基于分歧的半监督学习[J]. 自动化学报,2013,39(11):1871-1878.
ZHOU Zhi-hua. Disagreement-based semi-supervised learning[J]. Acta Automatic Sinica, 2013, 39(11): 1871-1878. (in Chinese)

[3] 蔡毅,朱秀芳,孙章丽,等. 半监督集成学习综述[J]. 计算机科学,2017,44(s1):7-13.
CAI Yi, ZHU Xiu-fang, SUN Zhang-li, et al. Semi-supervised and ensemble learning: a review[J]. Computer Science, 2017, 44(s1):7-13. (in Chinese)

[4] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[A]. Eleventh Conference on Computational Learning Theory[C]. New York: ACM Press, 1998. 92-100.

[5] ZHOU Z H, LI M. Tri-training: exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(11):1529-1541.

[6] LI M, ZHOU Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2007, 37(6):1088-1098.

[7] 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.

[8] 于重重,商利利,谭励,等. 一种增强差异性的半监督协同分类算法[J]. 电子学报,2013,41(1):35-41.
YU Chong-chong, SHANG Li-li, TAN Li, et al. A semi-supervised collaboration classification algorithm with enhanced difference[J]. Acta Electronica Sinica, 2013, 41(1):35-41. (in Chinese)

[9] 陈思,苏松志,李绍滋,等. 基于在线半监督boosting的协同训练目标跟踪算法[J]. 电子与信息学报,2014,36(4):888-895.
CHEN Si, SU Song-zhi, LI Shao-zi, et al. A novel co-training object tracking algorithm based on online semi-supervised boosting[J]. Journal of Electronics & Information Technology, 2014, 36(4):888-895. (in Chinese)

[10] 张磊,邵振峰,周熙然,等. 聚类特征和SVM组合的高光谱影像半监督协同分类[J]. 测绘学报,2014,43(8):

855 – 861.
ZHANG Lei, SHAO Zhen-feng, ZHOU Xi-ran, et al. Semi-supervised collaboration classification for hyperspectral remote sensing image with combination of cluster feature and SVM[J]. Acta Geodaetica et Cartographica Sinica,2014,43(8):855 – 861. (in Chinese)

[11] 邹保平,戚伟强. 基于改进 Co-Forest 的主机故障预警方法[J]. 电子设计工程,2017,25(5):65 – 69.
ZOU Bao-ping, QI Wei-qiang. Fault alarming method for host hardware based on improved Co-Forest[J]. Electronic Design Engineering, 2017, 25(5): 65 – 69. (in Chinese)

[12] JIA P, HUANG T, DUAN S, et al. A novel semi-supervised electronic nose learning technique: M-Training[J]. Sensors,2016,16(3):370.

[13] 杨艺,韩德强,韩崇昭. 一种基于证据距离的多分类器差异性度量[J]. 航空学报,2012,33(6):1093 – 1099.
YANG Yi, HAN De-qiang, HAN Chong-zhao. A novel diversity measure of multiple classifier systems based on distance of evidence[J]. Acta Aeronautica et Astronautica Sinica,2012,33(6):1093 – 1099. (in Chinese)

[14] 韩德强,杨艺,韩崇昭. DS 证据理论研究进展及相关问题探讨[J]. 控制与决策,2014,29(1):1 – 11.
HAN De-qiang, YANG Yi, HAN Chong-zhao. Advances in DS evidence theory and related discussions[J]. Control and Decision,2014,29(1):1 – 11. (in Chinese)

[15] 杜利敏,徐扬. 基于证据理论的不平衡数据半监督分类方法[J]. 计算机应用研究,2018,35(2):342 – 345.
DU Li-min, XU Yang. Semi-supervised classification method for imbalanced data based on evidence theory[J]. Application Research of Computers,2018,35(2):342 – 345. (in Chinese)

[16] LIU Z, PAN Q, MERCIER G, et al. A new incomplete pattern classification method based on evidential reasoning [J]. IEEE Transactions on Cybernetics,2015,45(4):635 – 646.

[17] 杨新武,马壮,袁顺. 基于弱分类器调整的多分类 Adaboost 算法[J]. 电子与信息学报,2016,38(2):373 – 380.
YANG Xin-wu, MA Zhuang, YUAN Shun. Multi-class Adaboost algorithm based on the adjusted weak classifier [J]. Journal of Electronics & Information Technology, 2016,38(2):373 – 380. (in Chinese)

[18] 孙伟超,许爱强,李文海. 区间信度结构下的证据合成方法研究[J]. 电子学报,2016,44(11):2726 – 2734.
SUN Wei-chao, XU Ai-qiang, LI Wen-hai. Approaches for combination of interval-valued belief structures[J]. Acta Electronica Sinica, 2016, 44(11): 2726 – 2734. (in Chinese)

[19] 郭强,何友,关欣,等. 一种多子焦元信度赋值非零情况下的 DSmT 近似融合推理方法[J]. 电子学报,2015,43(10):2069 – 2075.
GUO Qiang, HE You, GUAN Xin, et al. An DSmT approximate reasoning method on the condition of non-zero multiple focal elements [J]. Acta Electronica Sinica, 2015,43(10):2069 – 2075. (in Chinese)

[20] SMETS P, KENNES R. The transferable belief model [J]. Artificial Intelligence,1994,66(2):191 – 234.

[21] UCI machine learning repository[DB]. http:// archive.ics.uci.edu/ml/datasets,2017-10-10.

[22] PLATT J C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods [J]. Advances in Large Margin Classifiers,2000,10(4): 61 – 74.

作者简介

盛　凯　男,1991 年出生,山东兰陵人,海军工程大学兵器工程学院博士研究生. 主要研究方向为轨迹数据挖掘、机器学习及复杂系统建模与仿真.
E-mail:shengkai0214@foxmail.com

刘　忠　男,1963 年出生,山东龙口人,海军工程大学兵器工程学院教授、博士生导师. 主要研究方向为系统工程、复杂系统建模与仿真、系统集成技术.
E-mail:liuzhong531@yahoo.cn

周德超　男,1972 年出生,山东荣成人,海军工程大学兵器工程学院副教授,硕士生导师. 主要研究方向为复杂系统建模与仿真、数据挖掘、人工智能.
E-mail:13397190531@189.cn