ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 5,Sep. 2017,pp. 931－938
DOI:10.16337/j.1004-9037.2017.05.010

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于重采样和集成选择的适用于实体识别的多分类器系统

周　星[1]　刁兴春[2]　曹建军[2]　李　鑫[2]　王芳潇[2]

(1.解放军理工大学指挥信息系统学院,南京,210007;2.南京电讯技术研究所,南京,210007)

摘　要:实体识别常利用分类器根据记录对的字段相似度向量将记录对分为匹配、不匹配和可能匹配,因此分类器的准确性与实体识别的准确性直接相关。为提高分类准确性,本文基于重采样和集成选择技术构建一个多分类器系统。充分利用实体识别的特点,在分类之前发现分类困难的样本,并使重采样比率在一个区间内变化,生成一组重采样样本;然后用重采样后的样本训练分类器构建一个并行多分类器系统,强调分类器之间的差异度和稀疏度,从该多分类器系统中选择最优分类器子集,即最优的重采样比率组合,分别用非线性规划和极值方法求解该集成选择模型。实验结果表明,本方法与现有的多分类器系统相比具有更高的准确性。

关键词:实体识别;多分类器系统;重采样;集成选择;差异度

中图分类号:TP311　　**文献标志码**:A

Multiple Classifier System for Entity Resolution Using Resampling and Ensemble Selection

Zhou Xing[1], Diao Xingchun[2], Cao Jianjun[2], Li Xin[2], Wang Fangxiao[2]

(1. School of Command Information System, PLA University of Science and Technology, Nanjing, 210007, China; 2. Nanjing Telecommunication Institute, Nanjing, 210007, China)

Abstract: Classifiers are often used in entity resolution to classify record pairs into matches, non-matches and possible matches based on field similarity vector, in which case, the performance of classifiers is directly related to the performance of entity resolution. To improve the accuracy of classifier, a multiple classifier system is constructed. We make full use of the characters of entity resolution to distinguish the ambiguous instances before classification, vary the resampling ratio to generate a group of resampled instances, and use the resampled instances to train classifiers for constructing a parallel multiple classifier system. Moreover, we emphasize on the diversity and sparsity between classifiers to select the best classifier subset, and use non-linear programming and extreme value to solute the ensemble selection problem, respectively. Empirical experiments show the proposed multiple classifier system is superior to the state-of-art ones in accuracy due to resampling and ensemble selection.

Key words: entity resolution; multiple classifier system; resampling; ensemble selection; diversity

引　　言

根据应用背景,实体识别又被称为相似重复记录检测、记录链接和参照消歧等,其主要任务是识别

基金项目:国家自然科学基金(61371196)资助项目;中国博士后科学基金(201003797)特别资助项目;解放军理工大学预研基金(20110604,41150301)资助项目。
收稿日期:2015-09-25;**修订日期**:2016-09-05

相同或不同数据中表示同一客观对象的不同实体[1]。它是保证实体身份完整性和改进数据质量的重要途径,在国土安全、犯罪检测和客户关系管理中得到了广泛的应用。Christen P 总结了通用的实体识别过程,将其分为 3 个步骤:首先是索引,即聚合可能重复的记录以提升识别效率;其次是记录对比较,即运用相似度函数计算各个字段的相似度并生成字段相似度向量;最后是相似度向量分类,即根据字段相似度向量将记录对分为匹配、不匹配和可能匹配,其中可能匹配对应分类困难的样本,识别难度较大,通常需要专家参与以进一步的分类[2]。当分类器用于字段相似度向量分类时,实体识别就变为一个典型的分类问题。比如,Cochinwala 利用分类和决策树(Classification and regression tree, CART)[3],Bilenko 利用支持向量机(Support vector machine, SVM)[4]对其进行分类。为进一步提高分类的准确性,可以采用多分类器系统。比如,Tejada 等利用决策树构建多分类器系统,检测分类困难的样本,并对易分错的样本进行人工标记以强化学习,最终提高了分类准确性[5]。本文研究实体识别的第 3 个步骤,即相似度向量分类,主要工作为构建一个多分类器系统以提高分类的准确性,从而提高实体识别的准确性。

1 相关工作

多分类器系统的构造过程包括两步:首先训练多个分量分类器,再用投票、加权投票或者 stacking 的形式组合分量分类器[6]。典型的多分类器系统包括 Bagging[7]、Boosting[8]和 AdaBoost[9,10],其中 Bagging 采用 Bootstrap 采样方法对原有数据集进行采样,生成自助数据集,用其训练分量分类器,并对各分量分类器的结果投票组合,各自助数据集彼此独立,是一个典型的并行多分类器系统。AdaBoos 通过增大分类困难样本的权重来提高其被分量分类器选择的概率,使得算法聚焦于分类困难的样本,从而提高分类准确性。南京大学的周志华教授证明,从多分类器系统中选择合适的分类器子集,可以提升多分类器系统的准确性和效率,并且从并行分类器系统比从串行分类器系统中选择分类器子集更好[6,11]。

众多集成选择的方法主要分为两类:(1)贪心局部搜索,如前向搜索或后向搜索[12];(2)全局搜索,如周志华等利用遗传算法进行的搜索[11]。分量分类器之间的差异度被认为对多分类器系统的效果影响较大。目前差异度的度量和评估没有公认的方法[13],大多数研究用启发式方法选择具有差异度的分量分类器。曹建军基于皮尔森相关系数度量分量分类器之间的差异度,即相关性越小的分量分类器之间的差异度越大,并建立多分类器系统模型,要求每一个分量分类器都具有最大的分类正确性以及与其他分量分类器之间最小的相关性,该模型取得了较好的分类准确性[14]。Rafal L 等基于正确分类和条件错误的概率分别定义分量分类器之间的竞争度和成对差异度,并分别以竞争性为目标函数、差异度为约束条件以及以差异度为目标函数、竞争性为约束条件分别建立基于优化模型竞争性和多样性的动态集成选择(Dynamic ensemble selection-competence and diversity, DES-CD)的进行集成选择。该优化模型是典型的组合优化问题,因此他用模拟退火算法进行求解,实验表明,DES 方法具有比单个最好(Single best)以及多数投票更好的效果[15]。俞杨等用确定性数学优化方法管理分量分类器的差异度,基于分类器的夹角定义多分类器系统的成对差异度,并提出了差异度正则机(Diversity regularized machine, DRM)集成选择模型,理论分析表明该差异度定义起到了通常统计机器学习中正则化的作用。DRM 是一个二次约束二次规划问题,因此他用交替优化进行求解,实验表明,DRM 具有比 Bagging 和 AdaBoost 更好的效果[16]。Li Nan 等基于分类器预测输出的内积定义分类器的成对差异度,并提出了差异度正则集成选择(Diversity regularized ensemble pruning, DREP),并用迭代算法求解。实验表明,相比 Reduce Error[17], Kappa[17], Complementary[18]和 Margin Distance[18]方法,DREP 在准确性和分类器个数上都达到了更好的效果[6]。Yin Xucheng 也利用差异度和稀疏度进行了集成选择的研究,他根据集成歧义(Ensemble ambiguity)定义了一个凸的差异度度量,并将集成选择问题转换为凸优化问题,通过求极值,得到分类器权重的闭合解,相比 DRM 和 DREP 的迭代求解方法,该方法效率更高,但是也可能陷入局部最优解[19]。本文构建一个并行的关注分类困难样本的多分类器系统,并对其进行集成选择,以提高多分类的准确性。

2　基于重采样和集成选择的多分类器系统

多分类器系统的构造过程如图 1 所示。首先根据一组重采样比率,对原始数据进行重采样,生成一组重采样样本,用重采样样本训练一组 SVM 分量分类器,再根据分量分类器的差异度和稀疏度标准选择分类器子集,并加权投票,形成最终的投票决策。

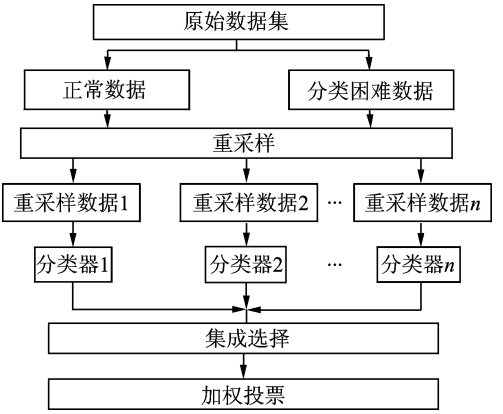

图 1　多分类器系统构造过程

Fig. 1　Outline of the construction of multiple classifier system

2.1　重采样

本文利用实体识别的特点,在分类前发现分类困难的样本,并采用重采样集成算法(Resampling ensemble algorithm,REA)[20]对分类困难的样本进行重采样。重采样比率对于重采样的效果影响较大,且通常最优重采样比率都为一个区间。与文献[20]求最优重采样比率不同,本文使重采样比率在一个区间内变化,得到一组重采样比率,生成一组重采样数据,用该重采样数据训练分量分类器,得到一个聚焦于分类困难样本的并行多分类器系统。实体识别的特点可以帮助在分类前发现分类困难的样本:实体识别通常需要计算记录对的相似度,并认为相似度高的记录对为匹配的可能性高,相似度低的记录对为不匹配的可能性高。而相似度不高不低的样本则可能为匹配,也可能为不匹配,它的分类较为困难,因此被称为分类困难的样本。

发现分类困难的样本的过程为:由于记录相似度呈近似正态分布[21],即匹配记录的记录相似度呈正态分布,不匹配记录的记录相似度也呈正态分布,而绝大部分正态分布的值都在 $(\mu_x-3\sigma_x,\mu_x+3\sigma_x)$ 区间,其中 μ_x 为期望,σ_x 为方差。令 $\boldsymbol{M}$ 为匹配记录对的记录相似度向量,$\boldsymbol{U}$ 为不匹配记录对的记录相似度向量,E_M 和 E_U 分别为 $\boldsymbol{M}$ 和 $\boldsymbol{U}$ 对应的期望,V_M 和 V_U 分别为 $\boldsymbol{M}$ 和 $\boldsymbol{U}$ 对应的方差。由于记录相似度呈近似正态分布,可以认为:相似度在 $(E_M-3V_M,1)$ 区间的记录对很可能为匹配,$(0,E_U+3V_U)$ 区间的记录对很可能为不匹配,相似度在 (E_U+3V_U,E_M-3V_M) 区间的记录对可能为匹配,也可能为不匹配,因此可以认为它们分类困难。基于该假设,给出如下的重采样算法。

算法 1　重采样算法

(1)输入:待重采样的数据 D:D 中每一个样本都是记录对的字段相似度向量,样本的类标指示该记录对是否匹配;重采样比例 r;

(2)初始化:$A=[]$,$N=[]$,$DO=[]$。将数据分为分类困难的数据和正常数据;获取样本个数 N;将数据分为匹配数据集 DM 和不匹配数据集 DU;对各样本(即记录对的字段相似度向量)取均值,以得到记录相似度向量 $\boldsymbol{S}_M$ 和 $\boldsymbol{S}_U$,计算 $\boldsymbol{S}_M$ 和 $\boldsymbol{S}_U$ 的期望和方差 E_M,E_U 以及 V_M,V_U;计算 DM 的下边界 LB,$LB=E_M-3V_M$,以及 DU 的上边界 U_B,$U_B=E_U+3V_U$

For $i=1,\cdots,N$

If $UB\leqslant S_i\leqslant LB$　　// S_i 是 D_i 的相似度

$A=A\cup\{D_i\}$　　// 分类困难样本

Else

$N=N\cup\{D_i\}$　　// 正常样本

(3)重采样

For $i=1,\cdots,\mathrm{round}(r\times N)$　　// round 为取整函数

从 A 中随机选择一个样本 A_r

$DO=DO\cup\{A_r\}$

For $j=1,\cdots,N-\mathrm{round}(r\times N)$

从 N 中随机选择一个样本 N_r

$DO=DO\bigcup\{N_r\}$

按照随机顺序重新排列 DO 中的各样本

(4)输出:重采样后的数据 DO。

算法 1 中,首先计算匹配记录对和不匹配记录对的记录相似度,再求记录相似度的均值和方差以确定分类困难样本相似度的上下边界,根据该边界将样本划分为正常样本和分类困难的样本。在重采样过程中,根据重采样比率,对分类困难的样本过采样,对正常样本欠采样,最终生成重采样的数据。由于初始化阶段的求期望、方差、上下边界以及划分复杂度为 $O(N)$,重采样过程为线性,复杂度也为 $O(N)$,因此,算法 1 的复杂度为 $O(N)$。

2.2 集成选择

2.2.1 集成选择模型

由于差异度和稀疏度在集成选择中应用广泛,本文也采用其作为集成选择的约束条件。根据文献[19],考虑差异度和稀疏度的通用集成选择模型为

$$\min_{w} f_{\text{loss}}(\boldsymbol{w})$$
$$\text{s.t.}\quad \text{sparsity}(\boldsymbol{w})\leqslant t_1,\text{diversity}(\boldsymbol{w})\geqslant t_2 \tag{1}$$

式中:t_1 为稀疏度控制参数,t_2 为差异度控制参数,$\boldsymbol{w}=(w_1,w_2,\cdots,w_n)$为分类器的权重向量。

损失函数的度量较多,比如最小均方误差,hinge 损失等。由于最小均方误差适用范围广泛、计算简单,因此本文采用其作为损失函数,定义为

$$f_{\text{loss}}(\boldsymbol{w})=\sum_{j=1}^{m}\frac{1}{2}(\boldsymbol{w}\boldsymbol{h}_j-y_j)^2 \tag{2}$$

式中:$\boldsymbol{h}_j$ 为所有分类器对第 j 个样本的预测输出向量,$\boldsymbol{Y}=(y_1,y_2,\cdots,y_m)$为测试样本的真实类标向量。

当分类是类标为 1 和 −1 的二分类时

$$f_{\text{loss}}(\boldsymbol{w})=\sum_{j=1}^{m}\frac{1}{2}(\boldsymbol{w}\boldsymbol{h}_j-y_j)^2\leqslant\sum_{j=1}^{m}\frac{1}{2}\times 2^2=2m \tag{3}$$

因此,可用 $2m$ 对损失函数进行归一化。

式(2)中的损失函数归一化后可以进一步写为

$$f_{\text{loss}}(w)=\frac{1}{4m}(\boldsymbol{w}\boldsymbol{H}-\boldsymbol{Y})(\boldsymbol{w}\boldsymbol{H}-\boldsymbol{Y})^{\mathrm{T}} \tag{4}$$

式中:$\boldsymbol{H}$ 为多分类器系统的输出矩阵,大小为 $n\times m$,其中 n 为分类器的个数,m 为样本的个数。

由于大部分差异度的度量都如文献[6,15,16]一样采用成对定义,且文献[6,15]中的定义已被证明能起到通常统计机器学习中的规范化的作用,因此本文也采用文献[6]中用到的差异度度量,其定义为

$$\text{div}(\boldsymbol{H})=1-\frac{1}{\sum_{1\leqslant i\neq j\leqslant n}1}\sum_{1\leqslant i\neq j\leqslant n}\text{diff}(\boldsymbol{h}_i,\boldsymbol{h}_j) \tag{5}$$

式中 diff 函数定义为

$$\text{diff}(\boldsymbol{h}_i,\boldsymbol{h}_j)=\frac{1}{m}\sum_{k=1}^{m}\boldsymbol{h}_i(\boldsymbol{x}_k)\boldsymbol{h}_j(\boldsymbol{x}_k)=\frac{1}{m}\boldsymbol{h}_i\boldsymbol{h}_j^{\mathrm{T}} \tag{6}$$

式中:$\boldsymbol{h}_i(\boldsymbol{x}_k)$ 是第 i 个分类器对样本 $\boldsymbol{x}_k$ 的预测。

当分类是类标为 1 和 −1 的二分类时

$$\text{div}(\boldsymbol{H})=1-\frac{1}{\sum_{1\leqslant i\neq j\leqslant n}1}\sum_{1\leqslant i\neq j\leqslant n}\frac{1}{m}\boldsymbol{h}_i\boldsymbol{h}_j^{\mathrm{T}}=$$
$$1-\frac{1}{mn(n-1)}(1,\cdots,1)\begin{pmatrix}0 & \boldsymbol{h}_1\boldsymbol{h}_2^{\mathrm{T}} & & \boldsymbol{h}_1\boldsymbol{h}_n^{\mathrm{T}}\\ \boldsymbol{h}_2\boldsymbol{h}_1^{\mathrm{T}} & 0 & & \boldsymbol{h}_2\boldsymbol{h}_n^{\mathrm{T}}\\ & & \ddots & \\ \boldsymbol{h}_n\boldsymbol{h}_1^{\mathrm{T}} & \boldsymbol{h}_n\boldsymbol{h}_2^{\mathrm{T}} & & 0\end{pmatrix}(1,\cdots,1)^{\mathrm{T}} \tag{7}$$

由于

$$\begin{pmatrix} 0 & \boldsymbol{h}_1\boldsymbol{h}_2^{\mathrm{T}} & & \boldsymbol{h}_1\boldsymbol{h}_n^{\mathrm{T}} \\ \boldsymbol{h}_2\boldsymbol{h}_1^{\mathrm{T}} & 0 & & \boldsymbol{h}_2\boldsymbol{h}_n^{\mathrm{T}} \\ & & \ddots & \\ \boldsymbol{h}_n\boldsymbol{h}_1^{\mathrm{T}} & \boldsymbol{h}_n\boldsymbol{h}_2^{\mathrm{T}} & & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{T}} & \boldsymbol{h}_1\boldsymbol{h}_2^{\mathrm{T}} & & \boldsymbol{h}_1\boldsymbol{h}_n^{\mathrm{T}} \\ \boldsymbol{h}_2\boldsymbol{h}_1^{\mathrm{T}} & \boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{T}} & & \boldsymbol{h}_2\boldsymbol{h}_n^{\mathrm{T}} \\ & & \ddots & \\ \boldsymbol{h}_n\boldsymbol{h}_1^{\mathrm{T}} & \boldsymbol{h}_n\boldsymbol{h}_2^{\mathrm{T}} & & \boldsymbol{h}_n\boldsymbol{h}_n^{\mathrm{T}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{T}} & 0 & & 0 \\ 0 & \boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{T}} & & 0 \\ & & \ddots & \\ 0 & 0 & & \boldsymbol{h}_n\boldsymbol{h}_n^{\mathrm{T}} \end{pmatrix} \tag{8}$$

而 $\boldsymbol{h}_i\boldsymbol{h}_i^{\mathrm{T}}=m, i=1,2,\cdots,n$,因此,$\mathrm{div}(\boldsymbol{H})=1-\dfrac{1}{mn(n-1)}(\boldsymbol{D}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}-nm)$,其中 $\boldsymbol{D}$ 为 n 维全 1 的行向量。

至于稀疏度,直接令 $0\leqslant w_i\leqslant t_1, i=1,\cdots,n$。

2.2.2　模型的求解

针对模型的求解过程,给出非线性规划和极值两种方法。

(1) 非线性规划集成选择(Nonlinear programming ensemble selection,NLPES)

由于稀疏度使部分分类器的权重为 0,而权重为 0 的分类器不会被用于计算最终分类器子集的差异度,因此将权重向量作为参数加入到差异度的定义中,即有

$$\mathrm{div}(\boldsymbol{H},\boldsymbol{w})=1-\frac{1}{mn'(n'-1)}(\mathrm{sgn}(\boldsymbol{w})\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}(\mathrm{sgn}(\boldsymbol{w}))^{\mathrm{T}}-n'*m) \tag{9}$$

式中:sgn 为符号函数,n' 为权重大于 0 的分类器个数,且 $n'=\mathrm{sgn}(\boldsymbol{w})(\mathrm{sgn}(\boldsymbol{w}))^{\mathrm{T}}$。

式(1)将转换为

$$\min_{\boldsymbol{w}} f(\boldsymbol{w})=\frac{1}{4m}(\boldsymbol{w}\boldsymbol{H}-\boldsymbol{Y})(\boldsymbol{w}\boldsymbol{H}-\boldsymbol{Y})^{\mathrm{T}}+\alpha\frac{1}{mn(n-1)}(\mathrm{sgn}(\boldsymbol{w})\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}(\mathrm{sgn}(\boldsymbol{w}))^{\mathrm{T}}-n'*m) \tag{10}$$

$$\text{s.t.}\quad 0<\boldsymbol{w}<\beta,\ \|\boldsymbol{w}\|=1$$

式中 α 和 β 为两个控制参数。式(10)是一个典型的二次规划问题,可以用现有的优化工具包求解。

相比 DRM 和 DREP 的交替优化和迭代求解方法,其求解较为简单,且可能取到全局最优解。

(2) 极值集成选择(Extrema ensemble selection,EES)

为进一步优化求解,将式(10)转换为

$$\min_{\boldsymbol{w}} f(\boldsymbol{w})=\frac{1}{4m}(\boldsymbol{w}\boldsymbol{H}-\boldsymbol{Y})(\boldsymbol{w}\boldsymbol{H}-\boldsymbol{Y})^{\mathrm{T}}+\alpha\frac{1}{mn(n-1)}(\boldsymbol{w}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{w}^{\mathrm{T}}-m\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})+\beta\boldsymbol{w}$$

$$\text{s.t.}\qquad \boldsymbol{w}\geqslant 0 \tag{11}$$

当 $\boldsymbol{w}>0$ 时,$f(\boldsymbol{w})$ 对 $\boldsymbol{w}$ 求偏导,可得

$$\frac{\partial f(\boldsymbol{w})}{\partial \boldsymbol{w}}=\left(\frac{1}{2m}+\frac{2}{mn(n-1)}\alpha\right)\boldsymbol{w}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}+\beta-2\boldsymbol{Y}\boldsymbol{H}^{\mathrm{T}} \tag{12}$$

令式(12)等于 0,可得

$$\boldsymbol{w}=(2\boldsymbol{Y}\boldsymbol{H}^{\mathrm{T}}-\beta)\left(\frac{1}{2m}+\frac{2\alpha}{mn(n-1)}\right)^{-1}(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})^{-1} \tag{13}$$

当 $\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}$ 不可逆时,用 Moore-Penrose 伪逆替代 $(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})^{-1}$。

式(13)的结果将使得部分权重为 0,从而实现集成选择。相比于式(9),式(13)的求解更加简单,且能得到确定解,但是也可能陷入局部最优解。

3　实体识别效果评估实验

3.1　实验设置

实验采用 6 个模拟数据集和 4 个真实数据集。模拟数据集来源于 UCI 机器学习数据库,包括 Abalone,Dermatology,Innosphere,Breast cancer,Seismic 和 ILPD(为方便处理,仅选择数值型和标称型字段),并用一个重复生成工具生成重复数据。首先取前 round($ri\times NI$)个样本(round 为取整函数),生成重复数据,为每个样本随机选择 round($ra\times NA$)字段,再根据字段类型,替换字段取值。即随机改变样

本部分字段的取值，其余字段不变，以生成重复数据。真实数据集使用文献[22]中用到的 Abt_buy，Amazon_gp 和 Dblp_acm 以及文献[23]中用到的 Cora。

字段相似度计算时，对字符型数据，使用 Jaccard 相似度[24]，数值型数据使用 $s(a,b)=1-\frac{\|a|-|b\|}{\max(|a|,|b|)}$，枚举型数据使用 $s(a,b)=\begin{cases}1 & a=b\\0 & \text{其他}\end{cases}$。

对每组数据，都使用 5 重交叉验证，采用随机选取的 4/5 的数据作为训练数据，1/5 的数据作为测试数据，并把本方法与 Gentle AdaBoost[9]，Bagging[7]，DREP[16]和 REA[20]进行对比。对 Bagging，将训练数据进行 21 次 Bootstrap 采样，训练 21 个 SVM，运用多数投票得到结果，重复 10 次实验，取平均值和方差作为最终结果；对 DREP，首先运行一次 Bagging，用 Bagging 的结果作为 DREP 的输入，并通过权衡参数从 0.05 到 0.5 的变化，得到 10 个结果，取平均值和方差作为最终结果；REA 是对训练数据划分出分类困难的样本和正常样本，对其进行重采样并训练分类器，重采样比率采用文献[20]中提出的经验重采样比率公式计算，再用测试数据进行测试；Gentle AdaBoost 使用 Alexander Vezhnevets 的 MATLAB 代码实现，分别用训练数据和测试数据直接进行训练和测试；对 NLPRES，将训练数据按照 0.4～0.6，步长为 0.01 变化的重采样比率进行重采样，训练 SVM 后得到 21 个 SVM 组成的分类器系统，再根据式(8)按照 α 取值为 1，β 取值为 0.7 进行集成选择；对 ERES，其 α 和 β 的取值按照文献[18]中的网格搜索算法进行选择。SVM 采用径向基函数(Radial-basis function，RBF)核函数，RBF 的宽度取值为 0.4，折衷系数为 100。针对式(10)，用 MATLAB 优化工具箱中的 fmincon 函数进行求解。

3.2 实验结果及分析

准确性对比结果如表 1 所示，其中，对每个数据，某个方法的结果如果在显著性水平为 0.1 的 t 检验下，比 Bagging 明显地好(或坏)，则用‘●’(或者是‘○’)标记，各数据的最优结果加粗。表格的最后为 win/tie/loss 计数。

从表 1 可以看出，DREP，Gentle AdaBoost 和 REA 在 7 个数据集上都比 Bagging 好。Gentle AdaBoost 在 3 个数据集上达到了 100%的准确性，但是它不稳定，在 3 个数据集上比 Bagging 较差。相对而言，DREP 在 2 个数据上比 Bagging 差一点，但是差别不大，这一方面是由 DREP 的权衡参数引起的，另一方面由于 Bagging 的结果是 10 次实验的均值，而 DREP 仅选择了 1 次 Bagging 的结果作为输入。相比而言，REA 的准确性提升不明显。这可能是由于文献[20]中提出的重采样比率计算公式为针对不平衡数据，而不是针对分类困难数据。

表 1 各方法的分类准确率对比

Tab. 1 Comparison of classification accuracy for different methods

数据集	Bagging	DREP	Gentle AdaBoost	REA	NLPES	EES
Breast cancer	0.899 6±0.000 2	0.908 8±0.000 2●	0.896 9○	0.87○	**0.932 7●**	0.919 3●
Innosphere	0.964 7±0.000 1	0.973 2±0●	**1●**	0.946 4○	0.991 1●	0.982 1●
Dermatology	0.954 1±0.000 3	0.967 6±0.000 2●	0.967 2●	0.956 1●	**0.982 5●**	0.964 9●
ILPD	0.961 9±0.000 1	0.972 1±0●	**1●**	0.957 1○	0.992 9●	0.964 3●
Seismic	0.977 9±0	0.975 6±0○	0.996 1●	0.990 3●	0.994 2●	0.992 2●
Abalone	0.981 2±0	0.981 2±0.000 1	0.988 0●	0.988 0●	**0.994 0●**	**0.994 0●**
Abt_buy	0.941 7±0	0.944 4±0●	0.937 2○	0.9465●	**0.953 5●**	0.951 2●
Amazon_gp	0.966 9±0	0.964 0±0○	0.952 5○	0.972 9●	**0.979 6●**	0.970 6●
Dblp_acm	0.993 8±0	0.997 5±0●	**1●**	0.996 6●	0.996 6●	0.993 3
Cora	0.940 7±0.000 1	**0.965 0±0●**	0.955 0●	0.950 0●	0.960 0●	0.950 0●
win/tie/loss	—	7/1/2	7/0/3	7/0/3	10/0/0	9/1/0

相比于 Bagging，NLPES 和 EES 在各数据集上的准确性均得到了提高。相比其他方法而言，NLP-

ES 在 5 个数据集上取得了最高的准确性,在余下的 5 个数据上接近最高准确性,虽然 EES 准确性的提升不如 NLPES,但是它的准确性仍然至少不低于 Bagging,能够得到稳定的改善。而 Bagging,Drep,REA 和 Gentle AdaBoost 都有比 Bagging 差的情况。综上,相比单纯的多分类器系统(Bagging)、单纯的集成选择(DREP)或者单纯的重采样(REA),本文的方法由于结合了重采样和集成选择,可以达到较高的准确性,且表现稳定。不同方法选择的分类器个数对情况如表 2 所示。从表 2 可以看出,与 NLPES 和 EES 相比,DREP 分类器的个数更多。可能的解释是重采样过程引入的数据使得分量分类器之间的差异度增加,因此有必要保留差异度高的分量分类器;另一个可能的解释是集成选择中参数的影响。相比而言,NLPES 在准确性和分类器个数上都优于 EES,然而 EES 效率更高。

表 2　各方法选择的分类器个数对比

Tab. 2　Comparison of selected classifier numbers

数据集	DREP	NLPES	EES	数据集	DREP	NLRES	EES
Breast cancer	4.6	7	11	Abalone	2.0	7	11
Innosphere	1.0	2	7	Abt_buy	3.8	6	12
Dermatology	2.8	5	7	Amazon_gp	4.9	5	7
ILPD	4.0	8	18	Dblp_acm	2.8	8	3
Seismic	3.7	7	9	Cora	4.3	5	9

由于重采样的时间复杂度较小,本文方法与 Bagging 相比,在多分类器系统构建时效率没有明显降低。并且,相比 DRM 和 DREP 的交替优化和迭代方法,本文的非线性规划和极值方法的求解过程也更加简单。

4　结束语

本文通过对分类困难的样本进行重采样,并使重采样比率在一个区间变化,生成一组重采样样本,用重采样样本训练一组 SVM 分类器,构建多分类器系统,再利用差异度和稀疏度进行集成选择,选择最优分类器子集,也即是最优重采样比率组合。然后采用非线性规划方法和极值方法进行求解。实验结果表明,相比单纯的多分类器系统、单纯的集成选择或者单纯的重采样,本文方法由于结合了重采样和集成选择,可以达到较高的准确性。下一步需要进一步分析式(10)和式(13)中控制参数的影响,以及扩展更多的重采样区间。本文方法适用于可以预先确定分类困难样本的情况。

参考文献:

[1] Elmagarmid A K, Ipeirotis P G, Verykios V S. Duplicate record detection: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(1): 1-16.

[2] Christen P. A survey of indexing techniques for scalable record linkage and deduplication[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(5): 1537-1555.

[3] Cochinwala M, Kurien V, Lalk G, et al. Efficient data reconciliation[J]. Information Sciences, 2001, 137: 1-4.

[4] Bilenko M, Mooney R J, Cohen W W, et al. Adaptive name matching in information integration[J]. IEEE Intelligent System, 2003, 18(5): 16-23.

[5] Tejada S, Knoblock C A, Minton S. Learning domain-independent string transformation weights for high accuracy identification[C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, New York: ACM Press, 2002: 350-359.

[6] Li Nan, Yu Yang, Zhou Zhihua. Diversity regularized ensemble pruning[C]// Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Bristol, UK: Springer Berlin Heidelberg Press, 2012: 1-16.

[7] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(3): 123-140.

[8] Schapire R E. The strength of weak learnability[J]. Machine Learning, 1990, 5: 197-227.
[9] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting[J]. The Annals of Statistics, 2000, 38(2): 337-374.
[10] Vezhnevets A, Vezhnevets V. Modest AdaBoost-teaching AdaBoost to generalize better[J]. Graphicon, 2005, 12(5): 987-997.
[11] Zhou Zhihua, Wu Jianxin, Jiang Yuan, et al. Genetic algorithm based selective neural network ensemble[C]// Proceedings of the 17th International Joint Conference on Artificial Intelligence. Seattle, WA: Morgan Kaufmann Press, 2001, 2: 797-802.
[12] Partalas I, Tsoumakas G, Vlahavas I. A study on greedy algorithms for ensemble pruning[R]. Technical Report TR-LPIS-360-12. Greece: Department of Informatics, Aristotle University of Thessaloniki, 2012.
[13] Zhou Zhihua, Li Nan. Multi-information ensemble diversity[C]// Proceedings of the 9th International Workshop on Multiple Classifier System. Cairo, Egypt: Springer Berlin Heidelberg Press, 2010: 134-144.
[14] 曹建军. 基于提升小波包和改进蚁群算法的自行火炮在线诊断研究[D]. 石家庄: 军械工程学院, 2007.
Cao Jianjun. Research on on-line fault diagnosis for self-propelled gun based on lifting wavelet package and improved ant colony optimization[D]. Shijiazhuang: Ordnance Engineering College, 2007.
[15] Rafal L, Marek K, Tomasz W. Optimal selection of ensemble classifiers using measures of competence and diversity of base classifies[J]. Neurocomputing, 2014, 126: 29-35.
[16] Yu Yang, Li Yufeng, Zhou Zhihua. Diversity regularized machine[C]// Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain: AAAI Press, 2011: 1603-1608.
[17] Margineantu D, Dietterich T. Pruning adaptive boosting[C]// Proceedings of the 14th International Conference on Machine Learning. Nashville, TN: Morgan Kaufmann Press, 1997: 211-218.
[18] Martinez-Munoz G, Hernandez-Lobato D. Suarez A. An analysis of ensemble pruning techniques based on ordered aggregation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009: 31(2), 245-259.
[19] Yin Xucheng, Huang Kaizhu, Yang Chun, et al. Convex ensemble learning with sparsity and diversity[J]. Information Fushion, 2014, 20: 49-59.
[20] Qian Yun, Liang Yanchun, Li Mu, et al. A resampling ensemble algorithm for classification of imbalance problems[J]. Neurocomputing, 2014, 143: 57-67.
[21] Zhou Xing, Diao Xingchun, Cao Jianjun. A data cleaning switch technology based on cloud model[C]// International Conference on Information Quality. Xi′an, China: ACM Press, 2014.
[22] Papadakis G, Koutrika G, Palpanas T, et al. Meta-blocking: Taking entity resolution to the next level[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1946-1960.
[23] Calado P, Herschel M. Efficient and effective duplication detection in hierarchical data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(5): 1028-1041.
[24] Xiao Chuan, Wang Wei, Lin Xuemin, et al. Efficient similarity joins for near-duplicate detection[J]. ACM Transactions on Database Systems, 2011, 36(3): 15.

作者简介:

周星(1988-),男,博士研究生,研究方向:数据工程,E-mail:zx0327@163.com。

刁兴春(1962-),男,研究员,研究方向:数据工程。

曹建军(1975-),男,工程师,研究方向:信息质量、进化算法。

李鑫(1984-),男,工程师,研究方向:数据工程。

王芳潇(1979-),女,工程师,研究方向:数据工程。