基于集成分类的高维数据实体分辨*

刘 艺[1], 刁兴春[1], 曹建军[2], 尚玉玲[1]
(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: 针对高维数据实体识别问题,为了有效利用高维特征的富信息,提高分辨性能,提出一种随机组合集成分类器。定义基分类器的分类性能指标,将分类正确性和特征子集的个数作为设计基分类器两个目标,使用聚合函数将其转换为单目标优化问题。采用蚁群优化求解基分类器模型,提出利用最大信息系数度量特征的相关性作为蚁群优化启发式信息,使用谷元距离度量选择特征多样性差异最大的基分类器组合集成分类器,集成分类器的决策函数采用投票表决输出。在标准数据集上进行验证与对比,结果表明了该方法的有效性。

关键词: 实体分辨; 高维数据; 集成分类器; 蚁群优化; 最大信息系数
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2018)03-0689-05
doi:10.3969/j.issn.1001-3695.2018.03.011

High-dimensional data entity resolution based on ensemble classifying

Liu Yi[1], Diao Xingchun[1], Cao Jianjun[2], Shang Yuling[1]
(1. *College of Command Information Systems, PLA University of Science & Technology, Nanjing* 210007, *China*; 2. *Nanjing Telecommunication Technology Institute, Nanjing* 210007, *China*)

Abstract: In order to effectively use rich information to improve performance of entity resolution in high-dimensional data, this paper proposed a random combinational ensemble classifiers' model. It defined the base classifier's classification performance's indicators, used the classification success rate and feature's number as two objects for optimizing base classifier, and adopted an aggregation function to transform them into a single objective optimization problem. It applied ant colony optimization to design base classifier, and adopted maximal information coefficient to measure correlation between features as heuristic information. The ensemble classifiers were composed of base classifiers which had the best diversity evaluated by Tanimoto distance, and used voting way to decide the output of ensemble classifiers. This paper adopts some benchmark datasets to evaluate the method, and the results show the effectiveness of the method.

Key words: entity resolution; high-dimensional data; ensemble classifiers; ant colony optimization; maximal information coefficient

0 引言

实体分辨是数据清洗中最为常见和重要的问题。实体分辨又称为记录链接(record linkage)、数据匹配(data matching)、实体识别、记录去重(record deduplication)、共指消解(co-reference resolution)、姓名消歧(name disambiguation)等,它的任务是识别出描述同一客观实体的歧义表示[1]。

近年来,实体分辨得到了广泛的关注。文献[2]基于图提出 GHOST 框架,在定义新的相似性度量的基础上,利用倾向传播算法对联合作者的信息进行聚类,解决不同文献中作者的姓名消歧问题;由于数字图书馆中存在的大量相似重复记录问题,文献[3]提出了一种重复聚类的模型,该模型结合了非监督方法和监督方法,通过前者提高模型的灵活性和鲁棒性,后者则用来提高识别的准确性;文献[4]将置信度的概念引入到相似度的计算中,并提出一种自适应的基于规则的方法计算记录相似度;文献[5]研究了互联网中用户名的引用消歧问题,提出了一种仅依靠用户名特征进行用户身份同一性判定的方法,通过对用户身份同一性判定问题进行了形式化描述,将用户名特征分为直观特征和对比特征两类,并对用户名特征的概率分布进行了量化分析,实现用户名引用消歧;文献[6]提出了基于重采样和集成选择的实体分辨的多分类器系统,该系统对分辨困难的样本进行重采样,重采样时将重采样比率在一个区间内变化,以生成一系列重采样样本,并用重采样后的样本训练分类器以构建一个并行多分类器系统,再运用集成选择方法从该多分类器系统中选择最优分类器子集,也就是最优的重采样比率组合。

实体分辨最为常用的方法包括基于特征(属性)相似度(feature based similarity, FBS)、上下文(context-based methods)和基于关系的方法(relationship-based method),其中 FBS 的研究起步较早,是最基本的实体分辨方法。它是通过分类器比较记录各属性的相似程度,然后综合各属性的相似程度来度量记录的相似程度,判断记录是否匹配。

大数据时代,数据的高维性已经成为亟待解决的问题之一[7]。传统 FBS 方法对全部属性进行比较,在数据维度较高

收稿日期: 2016-11-01; **修回日期**: 2016-12-23 **基金项目**: 国家自然科学基金资助项目(61371196);中国博士后科学基金特别资助项目(201003797);解放军理工大学预研基金资助项目(20110604,41150301)

作者简介:刘艺(1990-),男,博士研究生,主要研究方向为数据质量、进化算法(albertliu20th@163.com);刁兴春(1964-),男,研究员,硕士,主要研究方向为数据工程;曹建军(1975-),男,工程师,博士,主要研究方向为数据质量、进化算法;尚玉玲(1990-),女,硕士研究生,主要研究方向为数据质量、进化算法.

时会降低系统的可用性,而且也易引入噪声数据,同时采用全部属性并不总能提高分类准确性。虽然有相关文献使用特征选择来进行降维,但在高维数据中也不可避免地损失了丰富的信息。针对此类问题,本文基于 FBS 方法,提出使用互补特征组合训练基分类器,在定义分类器分类性能的基础上,使用蚁群优化求解特征组合,使用最大信息系数初始化蚁群优化算法的启发式信息,使用谷元距离度量选择特征区别度较高的基分类器组成性能互补的集成分类器,最终将每个基分类器的结果通过"Max-Wins" voting 投票表决输出。采用标准数据集进行测试并与其对比。

1 问题描述

实体分辨是根据两条记录的相似程度判断它们是否为同一客观实体的过程,检测过程可以看成一个二分类过程,待分类别为匹配和不匹配两类,分别称为第一类和第二类。a)通过计算两记录对应属性值的相似特征值,得到代表两记录相似程度的相似特征向量;b)将向量输入 SVM 二类分类器,完成实体分辨。检测流程如图 1 所示[8]。

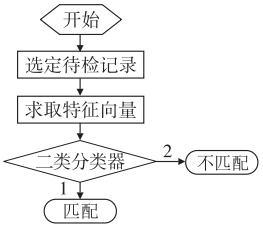

图1 实体分辨流程

为讨论方便,只讨论单张表的情况,设共有 n 个属性,属性集记为 $A=\{a_1,a_2,\cdots,a_n\}$,记第 i 条记录的第 k 个属性值为 $v_{ki},k=1,2,\cdots,n$,则第 i 条记录可记为 $r_i=(v_{1i},v_{2i},\cdots,v_{ni})$,再记 r_i 和 r_j 的第 k 个属性的相似特征值为 $s_{kij}=f_k(v_{ki},v_{kj})=f_k(v_{kj},v_{ki})$,可以得到 r_i 和 r_j 的相似特征向量 $\boldsymbol{V}_{ij}=(s_{1ij},s_{2ij},\cdots,s_{nij})$,将 $\boldsymbol{V}_{ij}$输入图 1 中的分类器,根据分类器的输出完成记录 r_i 和 r_j 是否匹配的检测。

2 分类器的分类性能度量与最大信息系数

2.1 分类器的分类性能度量

下面定义高维数据实体分辨的二类分类器性能度量指标。

分类正确率(classification success rate)P 为

$$P=\frac{\text{匹配的样本数}}{\text{参加测试的样本总数}}\times 100\% \tag{1}$$

虚警率(false alarm rate)R_{fa}为

$$R_{fa}=\frac{\text{匹配被误分为不匹配的样本数}}{\text{参加测试的匹配样本总数}}\times 100\% \tag{2}$$

漏诊率(fault not be recognized rate)R_{fn}为

$$R_{fn}=\frac{\text{不匹配误分为匹配的样本数}}{\text{参加测试的不匹配样本总数}}\times 100\% \tag{3}$$

定义二类分类器分类结果的分布矩阵为

$$\boldsymbol{p}=[p_{ii'}],i,i'=1,2 \tag{4}$$

其中:

$$p_{ii'}=\frac{\text{第 } i \text{ 类被分为第 } i' \text{类的样本数}}{\text{参加测试的第 } i \text{ 类样本数}}\times 100\% \tag{5}$$

其中:$p_{ii}(i=1,2)$为第 i 类样本的分类正确率,且其分类正确率 p_{ii}可计算为

$$p_{ii}=1-p_{ii'} \tag{6}$$

则分类正确率 P 为

$$P=P_ip_{ii}+P_{i'}p_{i'i'} \tag{7}$$

其中:P_i 为第 i 类样本的先验概率,对给定的测试样本集 P_i 可计算为

$$P_i=\frac{N_i}{N_i+N_{i'}} \tag{8}$$

其中:N_i 为第 i 类状态的样本数;$N_{i'}$为第 i'类状态的样本数。

同时,虚警率 R_{fa}计算(1 为匹配,2 为不匹配)为

$$R_{fa}=p_{12} \tag{9}$$

漏诊率 R_{fn}计算(1 为匹配,2 为不匹配)为

$$R_{fn}=p_{21} \tag{10}$$

P、R_{fa}和 R_{fn}之间存在式(11)的关系

$$1-P=P_1R_{fa}+(1-P_1)R_{fn} \tag{11}$$

从 P、R_{fa}和 R_{fn}的定义以及式(11)可以看出,R_{fa}和 R_{fn}是一对互相矛盾的指标,R_{fa}高 R_{fn}低,反之亦然,而 P 能够综合度量分类器的分类性能,因此采用 P 作为指标优化分类器的设计更为合适。

2.2 最大信息系数

在大数据时代,度量数据特征相关性的理论并不多见,因此通常许多研究都采用传统的相关性度量理论,如皮尔逊相关系数、斯皮尔曼相关系数等。然而它们都存在自身难以解决缺陷,皮尔逊相关系数对线性相关效果显著,但它不能度量线性关系的斜率以及非线性关系,也不能度量非函数关系;斯皮尔曼可以在一定程度上反映出非线性关系,但它的精度不高。

最大信息系数(MIC)的思想是:如果两个变量之间存在某种关系,那么在由这两个变量组成的散点图上可以按照某种方式画出一个网格,使得多数的点散布在该网格的几个单元格内[9]。

设考察的变量对样本构成有限集合 D,在集合 D 的散点图上,将元素按照 x 值划分到 x 个格子中,按 y 值划分到 y 个格子中,这种划分称为 $x\times y$ 划分。集合 D 的点散布在网格 G 上得到的概率分布记为 $D|_G$。集合 D 确定时,不同的网格 G 确定不同的概率分布。

定义 1 对有界集合 $D\subset \mathrm{R}^2$ 和正整数 x、y,有

$$I^*(D,x,y)=\max I(D|_G) \tag{12}$$

其中:$I(D|_G)$表示点集在网格 G 中分布 $D|_G$ 的互信息,最大值是在所有的 $x\times y$ 网格 G 中取最大值。

定义 2 二维数据集合 D 的特征矩阵元素为

$$\boldsymbol{M}(D)_{x,y}=\frac{I^*(D,x,y)}{\log\min\{x,y\}} \tag{13}$$

$\boldsymbol{M}(D)_{x,y}$是一个无穷矩阵。

定义 3 设二维数据集合 D 有 n 个样本,网格总数上限 $B(n)$,则其最大信息系数为

$$\mathrm{MIC}(D)=\max_{xy<B(n)}\{M(D)_{x,y}\} \tag{14}$$

MIC 的值在(0, 1)之间,它具有对称性,即 MIC(X, Y) = MIC(Y, X)。并且对于常数噪声函数关系和无噪声函数关系,MIC 依概率收敛到 1,而对于独立的两个变量,MIC 依概率收敛到 1。这也在很大程度上保证了 MIC 使用的广泛性和公平性。

3 基于特征选择的集成分类器设计

3.1 集成分类器模型设计

集成分类器能够在一定程度上提升分类的性能,且基分类器的多样性越强,分类结果的稳定性越好[10]。本节提出一种基于特征选择的集成分类器,基分类器综合考虑分类性能与特征规模,选择指定个数多样性最好的基分类器组合形成集成分类器,提高分类性能。

针对实体分辨问题,基于特征选择的二类分类器优化设计结构模型如下:

对规模为 L 的基分类器,记 P_l 为第 l 个分类器的分类正确率,q_l 为第 l 个分类器输入特征子集的基数,则第 l 个分类器输入特征子集及分类器的构造由如下目标函数确定:

$$\max\ P_l \tag{15}$$

$$\min\ q_l \tag{16}$$

即希望所设计的第 l 个分类器,同时具有最大的分类正确率,以及最小的特征子集规模。这是一个多目标问题,可以采用加权法将其转换为单目标问题。

基分类器设计完成后,从 L 个基分类器中随机选择规模为 M(奇数)的多样性差异最强的基分类器构成组合分类器,组合分类器的分类决策函数采用"Max-Wins" voting 投票表决:定义 f_{nl} 为第 n 个样本在第 l 个二类分类器上的决策函数:

$$f_{nl}=\begin{cases}1 & \text{匹配}\\ -1 & \text{不匹配}\end{cases} \tag{17}$$

则对个数为奇数 M 的基分类器,第 n 个样本的决策函数为

$$f(n)=f_{n1}\oplus f_{n2}\oplus\cdots\oplus f_{nM} \tag{18}$$

其中:⊕为异或操作。

3.2　集成分类器的蚁群算法实现

多目标问题通常并不存在各目标都为全局最优的解,而存在一非劣解集,称为 Pareto 最优解集,多目标优化的目的是力求找出一组解,尽可能全面地逼近 Pareto 解集,决策者可按需求进行评价,选出适用的满意解[11]。

一类基于元启发式算法全局搜索的次优子集求法得到了快速发展与应用,如蚁群算法、模拟退火算法、遗传算法、人工免疫算法、粒子群算法、蝙蝠算法、差分算法等[12]。

蚁群算法是一种新的元启发式算法,由于其具有很强的求解较好解的能力,较好的鲁棒性,信息正反馈,并行分布式计算及易于与其他启发式方法相结合等优点,在短期内得到了快速发展,应用领域也不断扩大,特别是在求解复杂多目标组合优化问题方面显示了其优越性[13]。本文使用蚁群算法来求解分类器设计模型。

为给出其蚁群算法实现,对 3.1 节的组合分类器设计模型进行如下分析:

a)为了能够减少模型求解的时间,需要限制特征选择的个数,即按照用户需求求解帕累托前沿的某一段,在多目标优化中称为融合用户偏好信息的优化[14]。对多类分类器而言,q_l 在 5 ~ 10 具有较好的运算效率和分类精度,但对于高维数据而言,可能需要更多的特征才可能得到较高的分类正确率,故将值的搜索限定在 1 ~ 20。

b)为了统一优化问题,需要将式(16)取倒数转换为最大化问题求解,将目标函数式(15)和转换后的式(16)用式(19)加权求和转换为单目标函数求解。

$$\max\left(\alpha_1 P_l+\alpha_2\frac{1}{q_l}\right) \tag{19}$$

其中:$\alpha_1>0,\alpha_2>0,\alpha_1+\alpha_2=1$。蚂蚁的全局最优解(特征子集)通过比较式(19)值来更新。

c)基分类器设计完成后,从中选择 M 个特征区分度较高的分类器构成组合分类器。对由 M 个特征子集组成的分类器系统$\{S_1,\cdots,S_M\}$,采用谷元距离度量基分类器之间的特征区分度,如式(20)所示。

$$T(S)=\frac{2}{M(M-1)}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\frac{|S_i\cap S_j|}{|S_i\cup S_j|} \tag{20}$$

当存在多个差异度相同的集成分类器系统时,随机选择其中一个集成分类器成为最终解,集成分类器的分类决策函数采用"Max-Wins" voting 投票表决。

3.3　蚁群算法设计

由 3.2 节的分析,求解多目标优化模型式(15)(16),关键是求解单目标优化模型式(19),而后者是典型的单目标无序组合优化问题,即子集问题。由于蚁群算法具有信息正反馈的属性,并且是一种"优化时学习"的算法[15],蚁群算法在子集问题上与其他进化算法相比具有优势,所以选择蚁群算法求解模型式(19)。文献[16]提出了一种求解子集问题的基于图的蚂蚁系统。针对子集问题,定义了构造图和等效路径,提出了基于等效路径增强的信息素更新策略。将信息素更新分为三种情况:本次迭代最优更新、变异更新和本次迭代不更新,兼顾算法的收敛速度和搜索能力。并以多维背包问题为例验证了系统的有效性和优越性。本文将此蚁群算法用于求解模型式(19)的子集问题。蚁群算法中的启发式因子使用特征的最大信息系数计算。根据实体分辨任务,将重复记录对分为相似重复和不相似重复两个类别(相似类别标记为 1,不相似类别标记为 2),设每个类别有 m 个样本,样本的类别已知,每个样本的特征规模为 k 个。

$$\{x_1^1,\cdots,x_m^1;x_1^2,\cdots,x_m^2\}$$

其中:$x_i^1=(x_{i,1},\cdots,x_{i,k})'$ 表示第 1 个类别第 i 个样本的 k 个特征组成的向量,第 2 个类别的特征向量与此类似。

根据问题特点,蚁群优化算法需要从 k 个特征中优先选择有效判别率大的相似特征,蚁群优化启发式信息求解算法的具体步骤如下所示。蚁群算法的其他部分参见文献[16]。

a)从类别 1 中随机选择一个样本作为对比样本 y^1,其他作为训练样本。

b)将对比样本的第 i 个特征与类别 1 的其他 $m-1$ 各训练样本的第 i 个特征一一对应,两个类共组成特征 i 的 $2(m-1)$ 个特征变量对。

c)计算步骤 b)中变量对的 MIC 值,共有 k 个 MIC 值,将其作为蚁群优化算法的启发式信息。

3.4　模型求解算法描述

根据以上分析,第 l 个分类器优化设计的具体算法描述如下:

```
begin
  初始化
  for num = 1:L
  {
  以式(19)为目标函数;
  按文献[16]的蚁群算法进行搜索;
  按以上分析 b)进行最优解更新
  if (P_num = 1, q_num = 1)
  循环终止
  }
  输出最优解
end
```

4　实验仿真与分析

4.1　实验数据及比较算法

实验数据来源于文献[17]中的 warpPIE10P,该数据集共有 10 个类,每个类含有 21 个实例,每个实例的特征维数是 2 420。每个类中取 20 对相似重复记录,不同类之间取 200 对

不相似重复记录组成共 400 组数据,其中 80% 的数据用做训练数据,20% 的数据作为测试数据。实验采用两种对比算法进行效果分析,文献[8]提出基于单个 SVM 的特征选择方法作为算法 1,将文献[18]提出的基于差分进化算法的特征选择方法作为算法 2,本文所提出的方法作为算法 3。

4.2 参数敏感性分析

本节对模型式(19)的参数敏感性作详细分析,为选择较好的运行参数提供依据。由模型式(19)可知,参数 $\alpha_1 > 0$, $\alpha_2 > 0$,且 $\alpha_1 + \alpha_2 = 1$,因此只对参数 α_1 作分析即可。将 α_1 在[0,1] 进行划分,取 11 组数值作为实验参数,即 $\alpha_1 = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$,在确定 α_1 的前提下,训练 10 个基分类器,按照 3.2 节分析 c)从中选择 3、5、7、9 个基分类器形成集成分类器,对这四种集成分类器的分类正确率的结果取均值作为 α_1 在当前取值下集成分类器的分类性能指标。仿真实验的结果如图 2 所示。

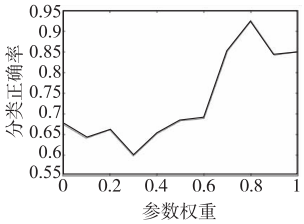

图2 参数权重敏感性分析图

从图 2 中可以看出,当 $0 < \alpha_1 < 0.3$ 时,分类正确率呈下降趋势,这是由于特征个数权重较高,导致选择的特征在分类效果上并不显著,特征训练的集成分类器在分类上存在欠拟合的现象,分类性能较弱。随着 α_1 的增加,选择特征的分类性能权重逐步增加,从而使得集成分类器的分类性能得到显著的增强。当 $\alpha_1 > 0.8$ 时,集成分类器的分类正确率有所下降,这是因为特征个数权重的下降导致算法选择了较多的特征,这些特征训练的基分类器组合形成集成分类器时会对数据集产生过拟合的现象,从而导致分类正确率的降低。所以,参数 α_1 在 0.8 左右时,集成分类器的综合性能可以达到较好的效果。

4.3 仿真结果与分析

算法 2 在原文中采用了两种分类器模型,即 K 近邻和 SVM 分类器,为了使得结果比较具有可解释性,三种对比算法使用的分类器模型均为 SVM,SVM 参数设置与算法 1 的设置相同[8],径向基核函数 $\delta = 0.4$, $C = 100$。候选基分类器个数设置为 10,模型式(19)的参数 $\alpha_1 = 0.8$。

蚁群算法的参数设置如下:初始时刻信息素 $\tau_{ij}(0) = 100$,信息素量和启发式信息的重要程度系数 $\alpha = 1$, $\beta = 1$,信息素挥发系数 $\rho = 0.2$,信息素增量调节常数 $Q = 200$,蚂蚁规模数 $M = 20$,停止条件为 40 次迭代搜索。算法 1 的参数设置与文献[8]相同,算法 2 中其他相关参数设置与文献[18]相同。

采用分类正确率指标度量算法运行结果,部分典型仿真结果列于表 1(仅列举部分结果),特征子集为蚂蚁寻找到的训练基分类器的特征集合。

如表 1 所示,算法 3 与算法 1 相比,在分类正确性上有较大的提升,这是由于相比较单个分类器的算法,集成学习方式本身具有提高分类器正确性的能力;其次是因为算法 3 先通过特征选择降低了噪声以及不相关特征,然后采用互补的特征组合进一步提高了对高维数据蕴涵特征信息的利用,并用来训练基分类器,进一步提升了集成分类器的分辨能力。

算法 3 与算法 2 相比,在分类正确率上也具有一定程度的提高,这是因为采用差分进化算法的算法 2 并未有效利用特征中蕴涵的先验信息,算法 3 使用最大信息系数对特征之间的相关性进行度量并作为蚁群算法的启发式信息,提高了蚁群算法获得较好解的概率,使得蚁群算法搜索的特征子集训练得到的基分类器要优于差分进化算法的选择及训练结果。

比较算法 3 在不同个数基分类器组合条件下的分类正确性可以发现,基分类器个数的增加并不总能提升分类正确性,这是因为所选特征的增加在某种程度上引入了噪声所致。另一方面,多数基分类器特征子集个数也并没有超过限定的上限,这也验证了特征个数的增加并不会提高分类正确性的前提。

表 1 算法分类正确性对比

实验次数	算法 1			算法 2			算法 3	
	特征子集	分类正确率	基分类器个数	特征子集	分类正确率	基分类器个数	特征子集	分类正确率
1	{295,21,1913,566,2036,1607,729,300,1253,2135,2082,1198,1462}	0.8813	3	{2,17,100,524,713,818,850,851,901,967,1043,1172,1388,1447,1554,2197} {3,5,94,363,514,690,718,893,1081,1391,1407,1747,2058,2222} {445,643,758,801,943,1035,1083,1149,1173,1401,1700,1747,1860,2078,2140,2238}	0.7271	3	{1031,1043,1287,73,1082,539,618,459,340,127} {621,870,1599,1774,2317,878,1660,1088,1889,1808} {1393,628,340,1765,2159,2076,2218,2062,57,1985,566}	0.9735
2	{671,894,1244,370,352,2103,606,1593,390,1551}	0.8781	5	{3,44,57,346,695,805,942,1023,1174,1259,1406,1415,1502,1669,1777,1780,1919,2101,2122} {304,459,1186,1236,1388,1483,1910,1965 {10,17,57,99,503,675,795,1175,1318,1426,2248} {2,10,12,396,401,516,960,1160,1199,1214,1507,1692,1887} {63,74,169,406,449,797,992,95,1055,1075,1166,1203,1215,1320,1609,1621,1636,1858}	0.8750	5	{1653,1889,987,2232,1053,729,1470,2016,2133,2025,2127,2150,1177,2230,1634,147,2087,1176} {15,579,1198,2324,2042,1543,2055,1861,2069,1065,2178,818,2027,896} {2267,1720,2031,791,1847,1566,2337,1779,2016,10,2383,1253,368,520} {840,1915,299,1859,615,22,2329,839,1759,2066,1653,2155,376,1198,16,2187} {2266,839,1004,1634,2061,2123,133,1994,840,2153,2378,2025,562,22,1198,949}	0.9633
3	{1059,2171,656,2012,934,1603,1004,1400,1912,840,2420,2078,1666,618,1969,1089,2321,1750}	0.8634	7	{11,61,74,207,285,856,905,957,1203,1621,1626,1629,1636,1637,1687,1803} {2,29,57,235,340,498,951,967,1746,1747} {61,62,63,122,137,284,500,608,965,967,1205,1271,1452,2121,2165,2195,2311} {349,921,1064,1077,1147,1198,1339,1622,1680}	0.7479	7	{2189,10,2379,1255,157,678,2403,423,2391,1114} {1926,2270,4,617,2093,2309,2326,215,626,2092,173,964,2208,1146,1218,2327,358,1830,2070,2352} {2253,2309,869,2322,2208,411,1118,185,1486,351,894,103,2094,2312,2256,99,913,1828} {2385,932,2380,1560,355,2014,1253,	0.9812

续表 1

实验次数	算法 1			算法 2			算法 3	
	特征子集	分类正确率	基分类器个数	特征子集	分类正确率	基分类器个数	特征子集	分类正确率
4	{1713,1761,15,840,116,1131,249,1702,1632,936,2187}	0.9054	9	{63,144,228,396,450,513,884,939,990,1101,1175,1195,1418,1620,1649,1687,1705,1899,2354} {350,1198,1704,1803,1361,2072} {7,63,112,113,685,976,1014,1088,1319,1539,1620,1627,1696,1697,1891,1906,1919,2026,2087} {52,57,92,235,299,995,1078,1225,1866,1916,1918,2002,2085} {47,56,59,100,209,400,487,514,652,662,1059,1164,1272,1304,1441,1831,2171,2195} {74,150,273,891,1002,1050,1459,1626,1623,1736,1902,1944,2103,2158,2318} {447,501,833,980,993,1126,1143,1148,1159,1179,1522,1863,1968,2082} {64,337,365,396,410,411,500,617,781,863,1140,1150,1632,1680,2053,2178,2233,2247} {101,366,416,529,784,891,1176,1245,1253,1317,1318,1494,1539,1606,1841,2023,2134} {97,98,355,535,608,717,771,790,805,995,1008,1031,1195,1205,1279,1571,1608,1627,1796,2295} {13,61,449,1252,2244,2245} {5,47,64,146,305,369,452,598,717,1023,1194,1396,1421,1725,1831,1840,1962,2070,2231,2380}	0.9125	9	2056,1779,2419,674} {1555,741,44,2312,1253,1723,2202,1930,2382,1548,3,2207,483,1671;2189,10,2379,1255,157,678,2403,423,2391,1114} {1926,2270,4,617,2093,2309,2326,215,626,2092,173,964,2208,1146,1218,2327,358,1830,2070,2352} {2397,1864,1852,2381,2419,2153,2059,1308,877,895,2165,923,1428} {2179,895,1198,2078,2231,683,2210,2234,925,1471,236,1062,1920,768,2247,2420,1906,2348} {2419,1916,989,1913,2070,1473,2291,2413,1152,1779,2385,1198,1560,2287,2119} {518,244,379,2210,2420,2229,2396,1033,1340,1059,2338,2062,2413,767,1310,1308} {1472,2301,1004,2397,521,691,1855,2071,1457,2383,1967,2418,2420,1303,1086,2133,2023,1779,409} {1555,2124,1088,2323,1594,2419,2173,507,1198,1458,1471,1034,385,1221,2192,651} {1524,784,2419,1308,1198,2013,1458,764,1047,988,1779,410,2080,2183,2265,2287} {70,1811,2217,804,1059,1779,1856,2326,404,2420,1407,1961,2193,1060,1253,1964,1528,2085,908,566} {699,768,2326,729,2411,2300,2327,1779,1909,1555,2419,15,988}	0.9390

5　结束语

本文将实体分辨看成是二分类问题,提出一种基于特征选择的集成分类器模型,采用互补特征子集训练基分类器,建立特征选择多目标优化模型,通过聚合转换为单目标函数并用蚁群算法对模型求解,用谷元距离度量选择特征差异度最强的基分类器组合形成集成分类器,在标准测试数据上进行仿真验证和算法对比,可以得到如下结论:a)仿真实例得到满意的分类正确性,说明采用区别度较高的特征训练的基分类器能进一步利用高维数据蕴涵的特征信息;b)特征选择多目标优化模型,综合考虑了分类算法的分类正确性和特征规模,实现了分类的效率和效果的综合最优;c)使用最大信息系数度量特征的相关性并作为蚁群算法的启发式信息,使得算法能够有效利用特征所蕴涵的先验信息,提高算法获得较好特征子集的概率;d)组合分类器中基分类器的个数与分类准确性的提升并没有固定的规律关系,适当地提高基分类器的个数能够在一定程度上提高分类准确性,但是过多的规模会引入噪声特征,影响分类性能。虽然实验结果表明了本文提出算法模型的有效性,但对高维数据蕴涵富信息的使用还有进一步提升的空间,另外,如何尽量减少噪声特征的影响,这也是下一步的研究方向。

参考文献:

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

[2] Fan X, Wang J, Pu X, *et al*. On graph-based name disambiguation [J]. Journal of Data & Information Quality, 2011,2(2):1-23.

[3] Zhu Jia, Yang Yi, Xie Qing,*et al*. Robust hybrid name disambiguation framework for large databases[J]. Scientometrics, 2014,98(3):2255-2274.

[4] Gu Qi, Zhang Yan, Cao Jian, *et al*. A confidence-based entity resolution approach with incomplete information[C]//Proc of IEEE International Conference on Data Science and Advanced Analytics. 2014.

[5] 刘东, 吴泉源, 韩伟红, 等. 基于用户名特征的用户身份同一性判定方法[J]. 计算机学报, 2015,38(10):2028-2040.

[6] Zhou Xing, Diao Xingchun, Cao Jianjun. A high accurate multiple classifier system for entity resolution using resampling and ensemble selection[J]. Mathematical Problems in Engineering, 2015, 2015(2):1-6.

[7] 徐宗本. 特邀报告: 大数据分析与处理的共性基础与核心技术[C]//第三届 CCF 大数据学术会议. 2015.

[8] 曹建军, 刁兴春, 杜鹢, 等. 基于蚁群特征选择的相似重复记录分类检测[J]. 兵工学报, 2010,31(9):1222-1227.

[9] Zhang Yi, Jia Shili, Huang Haiyun,*et al*. A novel algorithm for the precise calculation of the maximal information coefficient[J]. Scientific Reports, 2014,4(4):6662-6662.

[10] Li Xuchun, Wang Lei, Sung E. AdaBoost with SVM-based component classifiers[J]. Engineering Applications of Artificial Intelligence, 2008,21(5):785-795.

[11] Li Bingdong, Li Jindong, Tang Ke, *et al*. Many-objective evolutionary algorithms: a survey[J]. ACM Computing Surveys, 2015,48(1):1-35.

[12] Yang Xinshe. Bat algorithm: literature review and applications[J]. Bio-Inspired Computation, 2013, 3(5):141-149.

[13] Liao Tianjin, Socha K, Montes M A de Oca, *et al*. Ant colony optimization for mixed-variable optimization problems[J]. IEEE Trans on Evolutionary Computation, 2014,18(4):503-518.

[14] Kim J H, Han J H, Kim Y H, *et al*. Preference-based solution selection algorithm for evolutionary multiobjective optimization[J]. IEEE Trans on Evolutionary Computation, 2012,16(1):20-34.

[15] Ke L J, Zhang Q F, Battiti R. Using ACO in MOEA/D for multiobjective combinatorial optimization[EB/OL]. (2011-07-13)[2016-12-17]. http://www.cswww.essex.ac.uk.

[16] 曹建军, 张培林, 王艳霞, 等. 一种求解子集问题的基于图的蚂蚁系统[J]. 系统仿真学报, 2008,20(22):6146-6150.

[17] Hassanzadeh O, Chiang F, Lee H C, *et al*. Framework for evaluating clustering algorithms in duplicate detection[J]. Proceedings of the VLDB Endowment, 2009,2(1):1282-1293.

[18] Khushaba R N, Al-Ani A, Al-Jumaily A. Feature subset selection using differential evolution and a statistical repair mechanism[J]. Expert Systems with Applications, 2011,38(9): 11515-11526.