基于近邻信息和 PSO 算法的集成特征选取

刘全金[1,2],赵志敏[1],李颖新[3,4],俞晓磊[5]
(1. 南京航空航天大学理学院,江苏南京 210016;2. 安庆师范学院物理与电气工程学院,安徽安庆 246011;
3. 北京经纬纺机新技术有限公司,北京 100176;4. 北京市轻纺机械机器视觉工程技术研究中心,北京 100176;
5. 江苏省标准化研究院,江苏南京 210029)

摘 要: 提出了一种新的 PSO 特征选取方法. 以粒子对应特征组合的同类近邻样本和异类近邻样本间的距离关系作为类别可分性和粒子适应度函数. 以适应度函数加权的群体历史最佳、粒子历史最佳和粒子邻域内最佳个体信息共同指导粒子运动方向,搜索类内紧密、类间分离的最佳特征组合;同时,利用加权集成方法对 PSO 特征选取方法进行集成,以提高特征选取方法的稳定性和鲁棒性. 在 5 个高维数据集上的特征选取实验结果表明集成 PSO 特征选取方法的有效性和可行性.

关键词: 特征选取; PSO; 集成方法; 分类
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2016)04-0995-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.04.034

Ensemble Feature Selection Method Based on Neighborhood Information and PSO Algorithm

LIU Quan-jin[1,2], ZHAO Zhi-min[1], LI Ying-xin[3,4], YU Xiao-lei[5]
(1. *College of Science, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu* 210016, *China*;
2. *Department of Physics, Anqing Normal College, Anqing, Anhui* 246011, *China*;
3. *Beijing Jingwei Textile Machinery New Technology Co Ltd, Beijing* 100176, *China*;
4. *Beijing Light Industry and Textile Machinery Engineering Research Center for Machine Vision, Beijing* 100176, *China*;
5. *Jiangsu Institute of Standardization, Nanjing, Jiangsu* 210029, *China*)

Abstract: A new PSO algorithm is proposed in this paper for feature selection. Distances within the same class and between different classes are used as the index for distinguishing different classes, and thus can be used to construct the fitness function of particles in PSO. The direction of particles for searching optimal features which can result in close intra-class distance and far inter-class distance is determined by the current best solution of the particle and the optimal individual in particle neighborhood, weighted by the fitness function. Meanwhile, the PSO algorithm is aggregated by the weighted voting method to improve its stability and robustness. The experiment results on 5 high dimensional datasets show that the ensemble PSO algorithm is effective and feasible.

Key words: feature selection; PSO (particle swarm optimization); ensemble method; classification

1 引言

利用机器学习技术对高维数据降维,使降维后的数据仍承载初始数据中有效信息[1,2]. 特征选取方法根据特征评价函数,从原始数据中选取代表原始数据特性的关键特征[3,4]. 特征选取方法分为特征权重选取方法和特征子集选取方法. 特征权重选取方法根据特征在不同类别样本中的分布选取分类特征[5],或基于分类模型推算特征对分类的影响进而选取分类特征[6]. 特征子集选取方法包括穷举法、启发法和随机法等. 高维数据集的穷举法特征选取是 NP 问题;启发法为次优搜索算法,在迭代的过程中产生维数递增或递减的特征子集,能得到近似最优解[7],如顺序后向选择法(Sequential Backward Selection, SBS);随机法基于随机性从

收稿日期:2014-10-24;修回日期:2014-12-16;责任编辑:孙瑶
基金项目:国家自然科学基金(No. 61475071, No. 61173068, No. 10172043);教育部博士点基金(No. 20093218110024);江苏省自然科学基金青年基金(No. BK20141032);国家质检总局科技项目(No. 2013QK194);安徽省自然科学基金(No. 1608085QF157)

特征空间中选取特征子集,主要包括基于进化思想的遗传算法和粒子群算法等优化算法.

Kennedy 等人[8]提出粒子群(Particle Swarm Optimization,PSO)算法. 粒子群粒子通过与个体历史最佳和群体历史最佳的信息交流实现种群进化[9]. 二进制 PSO 算法[10]被用于解决 0-1 整数规划问题,也适于特征选取问题研究. 文献[11,12]提出基于分类器的二进制 PSO 特征选取算法(简称 BPSO 方法),以粒子个体对应特征组合分类结果为适应度函数进行分类特征选取研究. 适应度函数的特征组合分类交叉校验需要较大的计算量,选取的特征组合亦可能过拟合于适应度函数所用的分类模型. 另外,仅考虑粒子群群体历史最佳的全局模型虽然算法收敛速度快,但算法易于早熟.

集成特征选取方法通过集成来自不同特征选取算法、不同样本组成或不同特征选取范围选取的特征,从中确定最佳分类特征[13]. Abeel[14]和 Saeys[15]通过实验证明集成方法有助于提高特征选取方法鲁棒性和稳定性.

本文提出基于近邻信息的 BPSO 特征选取方法(以下简称 NBPSO 方法). 基于同类近邻样本和异类近邻样本信息定义粒子个体对应特征组合的类别可分性和粒子适应度函数,以保证特征组合类内紧密、类间分离;同时,引入粒子近邻信息,提出以类别可分性加权的群体历史最佳位置、粒子个体历史最佳位置和粒子邻域内最佳位置共同指导粒子个体运动方向,在特征空间中搜索最佳特征组合,研究高维数据的特征降维问题. 另外,本文提出基于加权集成方法对并行 NBPSO 方法所选特征集成,以提高 NBPSO 方法的稳定性和鲁棒性.

在 5 个高维数据集上进行的特征选取实验结果表明 NBPSO 方法在特征组合分类、算法复杂度和稳定度等方面优于基于分类器的 BPSO 特征选取方法,并证实了加权集成方法对 NBPSO 方法的促进作用. 这说明本文所提特征选取方法的有效性和可行性.

2 基于 BPSO 算法的特征选取方法

PSO 算法中每个粒子代表解空间中一个解. 初始阶段,群体粒子随机分布于解空间;进化过程中,粒子基于适应度函数,根据个体和群体历史最佳粒子位置更新速度和位置,逐渐收敛于解空间的最佳位置.

设在 m 维解空间中,第 k 代群体第 i 个粒子 $\boldsymbol{X}_i^k=[\boldsymbol{X}_{i1}^k \cdots \boldsymbol{X}_{id}^k \cdots \boldsymbol{X}_{im}^k]^{\mathrm{T}}$;粒子速度 $\boldsymbol{V}_i^k=[\boldsymbol{V}_{i1}^k \cdots \boldsymbol{V}_{id}^k \cdots \boldsymbol{V}_{im}^k]^{\mathrm{T}}$;粒子的历史最佳 $\boldsymbol{P}_i=[\boldsymbol{P}_{i1} \cdots \boldsymbol{P}_{id} \cdots \boldsymbol{P}_{im}]^{\mathrm{T}}$;粒子群群体的历史最佳 $\boldsymbol{P}_g=[\boldsymbol{P}_{g1} \cdots \boldsymbol{P}_{gd} \cdots \boldsymbol{P}_{gm}]^{\mathrm{T}}$. 如式(1)所示,进化过程中根据粒子的速度分量 $\boldsymbol{V}_{id}^k$、位置分量 $\boldsymbol{X}_{id}^k$ 与个体和群体历史最佳的关系更新速度分量:

$$\boldsymbol{V}_{id}^{k+1}=\omega\boldsymbol{V}_{id}^k+c_1r_1(\boldsymbol{P}_{id}-\boldsymbol{X}_{id}^k)+c_2r_2(\boldsymbol{P}_{gd}-\boldsymbol{X}_{id}^k) \quad (1)$$

式中,速度惯性因子 $\omega>0$;速度权重 $c_1>0$ 和 $c_2>0$;随机参数 $r_1\in[0,1]$ 和 $r_2\in[0,1]$;以最大速度 $v_{\max}$ 和最小速度 $v_{\min}$ 限制粒子移动速度范围.

基于二进制 PSO 算法的 BPSO 特征选取方法将所有备选特征作为特征选取搜索空间,搜索空间中不同位置粒子对应的特征组合. 对于训练集 $\mathbf{TrainD}\in\mathbf{R}^{m\times n}$($n$ 个样本、m 个特征),在 m 维二值解空间中粒子位置分量 $\boldsymbol{X}_{id}^k\in\{0,1\}$ 决定第 d 个特征的选择状态,分量值 1 或 0 表征特征被选取或被舍弃. 用 S 型函数 $S(\boldsymbol{V}_{id}^{k+1})=1/(1+\mathrm{e}^{-\boldsymbol{V}_{id}^{k+1}})$ 将速度分量压缩至(0,1)之间,根据式(2)确定该粒子最新位置:

$$\boldsymbol{X}_{id}^{k+1}=\begin{cases}1, & S(\boldsymbol{V}_{id}^{k+1})\geq\mathrm{rand}\\ 0, & S(\boldsymbol{V}_{id}^{k+1})<\mathrm{rand}\end{cases} \quad (2)$$

式中,rand 是 0 到 1 之间的随机数值.

粒子适应度函数 $f(\boldsymbol{X}_i^k)$ 以粒子对应特征组合的类别可分性为主要参数,很多文献定义特征组合分类正确率为类别可分性,同时,将特征组合维数作为适应度函数另一个参数,使之向低维特征组合进化. 进化终止时,$\boldsymbol{P}_g$ 为特征解空间最佳解向量,其非零元素 $\boldsymbol{P}_{gd}$ 对应的特征组成分类特征集合.

3 基于近邻信息的集成 NBPSO 特征选取方法

3.1 NBPSO 特征选取方法

本文提出基于近邻信息的 BPSO 特征选取方法(简称 NBPSO 方法). 综合运用全局搜索和局部搜索优势,将粒子邻域内最佳个体粒子信息纳入 BPSO 特征选取方法的粒子速度更新中,并以粒子对特征组合在同类近邻样本和异类同类样本间的距离关系作为粒子适应度函数,通过粒子适应度函数施加对粒子运动方向的影响.

以粒子间的汉明距离为标准寻找粒子 $\boldsymbol{X}_i^k$ 近邻粒子,近邻粒子中适应度值最大的粒子称粒子邻域内最佳粒子 $\boldsymbol{N}_i^k$. 更新粒子速度 $\boldsymbol{V}_i^{k+1}$ 时引入 $\boldsymbol{N}_i^k$,利用 $\boldsymbol{P}_i$、$\boldsymbol{P}_g$ 及 $\boldsymbol{N}_i^k$ 的适应度函数值均衡粒子 $\boldsymbol{X}_i^k$ 与 $\boldsymbol{P}_i$、$\boldsymbol{P}_g$ 及 $\boldsymbol{N}_i^k$ 之间的关系. 粒子速度分量更新公式为:

$$\begin{aligned}\boldsymbol{V}_{id}^{k+1}=&\omega\boldsymbol{V}_{id}^k+\frac{1}{f(\boldsymbol{P}_i)+f(\boldsymbol{P}_g)+f(\boldsymbol{N}_i^k)}(c_1r_1f(\boldsymbol{P}_i)(\boldsymbol{P}_{id}-\boldsymbol{X}_{id}^k)\\&+c_2r_2f(\boldsymbol{P}_g)(\boldsymbol{P}_{gd}-\boldsymbol{X}_{id}^k)+c_3r_3f(\boldsymbol{N}_i^k)(\boldsymbol{N}_{id}^k-\boldsymbol{X}_{id}^k))\end{aligned} \quad (3)$$

粒子群进化过程中补充粒子邻域内最佳粒子信息,有助于 BPSO 算法利用粒子的局部信息提高全局搜索能力;用适应度函数加权粒子间“位置差异”,引导粒子向适应度函数更大的位置移动;对适应度函数值归

一化,既有助于均衡 $\boldsymbol{P}_i$、$\boldsymbol{P}_g$和 $\boldsymbol{N}_i^k$之间关系,又有助于平衡粒子之间"位置差异"与粒子速度之间的数值关系(见算法 1).

算法 1　NBPSO 特征选取方法

输入:训练集 **TrainD**,$k=0$;

Step 1:粒子群初始化. 随机生成 L 个粒子个体 $\{\boldsymbol{X}_1^1 \cdots \boldsymbol{X}_i^1 \cdots \boldsymbol{X}_L^1\}$和粒子速度 $\{\boldsymbol{V}_1^1 \cdots \boldsymbol{V}_i^1 \cdots \boldsymbol{V}_L^1\}$,设置 $\boldsymbol{v}_{\max}$、$\boldsymbol{v}_{\min}$、ω、c_1、c_2、nearMiss、nearHit、最大迭代次数 $k_{\max}$ 和适应度阈值 T_f;

Step 2:根据式(4)计算粒子个体适应度函数值,寻找初始粒子邻域内最佳粒子 $\boldsymbol{N}_i^k$,确定群体历史最佳 $\boldsymbol{P}_g$;

Step 3:$k=k+1$,根据式(3)更新粒子速度分量 $\boldsymbol{V}_{id}$;

Step 4:根据式(2)更新粒子位置分量 $\boldsymbol{X}_{id}$;

Step 5:根据式(4)更新粒子个体适应度函数值;

Step 6:更新粒子邻域内最佳 $\boldsymbol{N}_i^k$,粒子历史最佳 $\boldsymbol{P}_i$和群体历史最佳 $\boldsymbol{P}_g$;

Step 7:若 $k>k_{\max}$或 $f(\boldsymbol{P}_g)>T_f$,则停止迭代,否则返回 Step3;

Step 8:$\boldsymbol{P}_g$对应特征组成分类特征集合.

利用样本间的欧氏距离考察粒子 $\boldsymbol{X}_i^k$对应的特征组合在 TrainD 样本中的类别可分性. 基于特征组合以样本与异类近邻样本间欧氏距离度量其与异类样本间的差异性,以样本与同类近邻样本间的欧氏距离考察其与同类样本间的相似性. 与同类样本间的欧氏距离越小,同类样本分布越紧密;与异类样本间的欧氏距离越大,异类样本分布越分散. 定义粒子适应度函数:

$$f(\boldsymbol{X}_i^k) = \frac{\sum_{j=1}^{n}\left(\sum_{n=1}^{\text{nearMiss}} \| \boldsymbol{D}_j(\boldsymbol{X}_i^k) - \boldsymbol{D}_j^n(\boldsymbol{X}_i^k) \|\right)}{n \times \text{nearMiss}} - \frac{\sum_{j=1}^{n}\left(\sum_{n=1}^{\text{nearHit}} \| \boldsymbol{D}_j(\boldsymbol{X}_i^k) - \boldsymbol{D}_j^n(\boldsymbol{X}_i^k) \|\right)}{n \times \text{nearHit}} + \frac{1}{\text{Dim}(\boldsymbol{X}_i^k)} \tag{4}$$

式中,n 是 **TrainD** 的样本个数,$\boldsymbol{D}_j$是第 j 个样本,nearHit 表示离 $\boldsymbol{D}_j$最近的同类样本个数,nearMiss 表示离样本 $\boldsymbol{D}_j$ 最近的异类样本个数;$\| \boldsymbol{D}_j(\boldsymbol{X}_i^k) - \boldsymbol{D}_j^n(\boldsymbol{X}_i^k) \|$ 表示基于粒子 $\boldsymbol{X}_i^k$对应特征组合样本 $\boldsymbol{D}_j$与同类近邻样本 $\boldsymbol{D}_j^n$间的欧氏距离,$\| \boldsymbol{D}_j(\boldsymbol{X}_i^k) - \boldsymbol{D}_j^{\bar{n}}(\boldsymbol{X}_i^k) \|$ 表示基于粒子 $\boldsymbol{X}_i^k$对应特征组合样本 $\boldsymbol{D}_j$与异类近邻样本 $\boldsymbol{D}_j^{\bar{n}}$ 间的欧氏距离,$\text{Dim}(\boldsymbol{X}_i^k)$表示粒子 $\boldsymbol{X}_i^k$对应特征组合中的特征个数.

适应度函数以样本与异类近邻样本间平均距离和样本与同类近邻样本间平均距离的差值度量特征组合的类别可分性,旨在搜索类内紧密、类间分离的特征组合. 适应度函数兼顾特征组合的类别可分性与特征组合维数,有利于 NBPSO 方法选取类间差异性大、类内相似度高的低维分类特征集合.

计算样本间距复杂度比样本分类交叉校验复杂度低,故 NBPSO 方法的算法复杂度比基于分类结果构建适应度函数的 BPSO 方法的算法复杂度低,消耗时间少.

3.2　集成 NBPSO 特征选取方法

利用 NBPSO 方法进行集成特征选取研究(简称 ENBPSO 方法). 首先,多个 NBPSO 特征选取过程并行进行;然后,对所选特征做加权集成,通过特征的集成权值高低确定分类特征.

如图 1,在特征选取范围内并行进行 N 个 NBPSO 特征选取过程. 每个 NBPSO 过程基于随机产生的粒子群初始群体确定各自的特征搜索范围,通过粒子与群体历史最佳、个体邻域内最佳和个体历史最佳的信息交换,调整进化速度和方向,进而获取群体历史最佳粒子对应特征组合. 通过特征组合集成确定特征的集成权值,选取权值高的特征作为分类特征.

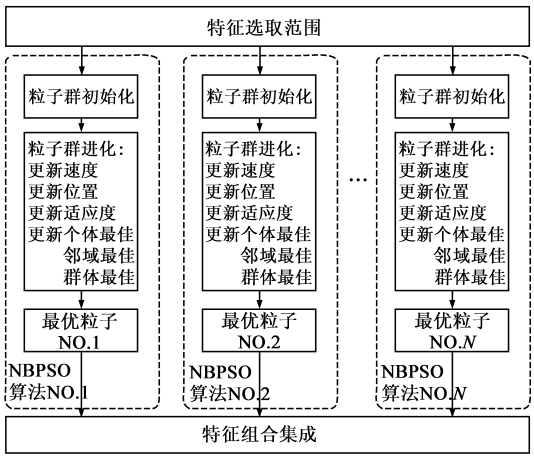


图1　集成NBPSO特征选取方法流程图

粒子群随机初始化使每个 NBPSO 过程的特征搜索范围各不相同,满足集成方法特征选取范围多样性条件,有利于提高特征选取方法鲁棒性和所选分类特征的分类性能.

粒子适应度函数式(4)前半段是粒子对应特征组合在训练集样本中异类样本之间差异性和同类样本之间相似性的差值,定义其为 NBPSO 过程最佳粒子对应特征组合 $\boldsymbol{S}_{\text{optimal}}$的类别可分性 $\text{Div}(\boldsymbol{S}_{\text{optimal}})$. 根据特征组合类别可分性对 N 次 NBPSO 所得特征组合进行加权集成,集成公式为:

$$\text{weight}_{\text{ensemble}} = \sum_{k=1}^{N}\left(\text{Div}(\boldsymbol{S}_{\text{optimal}}^k) \cdot \boldsymbol{S}_{\text{optimal}}^k\right) \tag{5}$$

以类别可分性指标加权特征组合的特征,集成后的特征权值能体现特征被选频次,又反映特征在异类样本和同类样本间的相对差异程度,权值高的特征对分类贡献大. 然而,如何根据特征权值确定选取阈值进而选取分

类特征集合目前尚无普适性标准.

利用顺序后向选取(SBS)方法,一次仅剔除一个特征形成嵌套的候选特征子集,然后根据特征子集评价函数从中找出寻找最佳特征集合. 这种方式选取效果较好,但对于高维数据特征选取问题,这种方式计算量太大,不易现实. 文献[6]提出一种折中办法,在递归特征剔除过程(RFE,Recursive Featrue Elimination)中,兼顾特征选取方法有效性和可行性,每次按比例选择一部分特征作为候选特征子集.

在 RFE 过程中用 ENBPSO 方法获取特征集成权值,基于特征集成权值生成候选特征子集,以候选特征子集分类结果为评价函数从中选取最佳特征集合(见算法 2).

算法 2 ENBPSO 特征选取方法

输入:$\boldsymbol{X}$ 训练集;$\boldsymbol{V}$ 校验集;$\boldsymbol{F}$ 候选特征子集;p % 特征剔除率;

Step 1:以 $\boldsymbol{X}$ 的特征为初始候选特征子集 $\boldsymbol{F}$;

Step 2:在 $\boldsymbol{X}$ 中分析、计算特征集成权值:

(1)基于 $\boldsymbol{F}$ 并行 N 次 NBPSO 特征选取,获得 N 个特征集合;

(2)根据式(5)集成特征权值;

Step 3:删除 $\boldsymbol{F}$ 中 p % 低权值特征 $\{\boldsymbol{f}\}$,生成新的候选特征子集 $\boldsymbol{F}=\boldsymbol{F}-\{\boldsymbol{f}\}$;

Step 4:基于 $\boldsymbol{F}$ 在 $\boldsymbol{X}$ 中训练 SVM(Support Vector Machine)和 KNN(K-Nearest Neighbor)分类器,并对 $\boldsymbol{V}$ 的样本分类;

Step 5:如果 $\boldsymbol{F}$ 的维数大于 3,则转至 Step 2,否则转至 Step 6;

Step 6:选取 AUC(Area Under roc Curve)值和正确率最高的 $\boldsymbol{F}$ 作为分类特征集合.

数据集被分为训练集和校验集. RFE 过程中,对每个新生成候选特征子集重新并行 N 次 NBPSO 特征选取,特征的动态集成权值能客观反映特征之间的相对重要性,有利于提高特征选取质量.

SVM(Support Vector Machine)分类器是 Vapnik 等人[6]基于结构风险最小化原理,运用统计学习理论推导出的分类模型. KNN(K-Nearest Neighbor)分类器是基于近邻思想演化而来的较为简单分类算法[4]. 分类正确率体现基于特征子集建立的分类器对样本的识别能力;分类 AUC(Area Under roc Curve)值表示 ROC(Receiver Operating Characteristics)曲线下面积[16],能客观反映异类样本数量不均匀时的分类结果.

4 特征选取实验

将 NBPSO 方法、ENBPSO 方法和基于 SVM 的 BPSO 方法以及 SVM-RFE 方法同时进行特征选取实验,通过实验比较 4 种方法在高维特征数据集上的特征选取性能. SVM-RFE 方法基于 SVM 分类模型在 RFE 过程中分析特征对分类模型的影响,进而选取分类特征集合[6]. 为不失一般性,在 5 个两类数据集上进行特征选取实验. 表 1 列出了 DLBCL、Acute Leukemia、Multiple Myeloma、Colon 和 Prostate 等 5 个基因表达谱数据集的相关参数.

表 1 基因表达谱数据集

基因表达谱数据集	基因数	样本数(正类/负类)	参考文献
DLBCL	7129	77(58/19)	[17]
Acute Leukemia	7129	72(25/47)	[18]
Multiple Myeloma	7129	105(31/74)	[19]
Colon	2000	62(22/40)	[20]
Prostate	12600	102(50/52)	[21]

4.1 特征选取实验参数设置

鉴于高维数据中存在大量与类别无关的噪声特征,在特征选取之前利用 Bhattacharyya 距离过滤噪声特征. Bhattacharyya 距离不但体现特征在不同类别样本间的差异,还反映特征在各类样本中的变化情况[22].

为客观比较 4 种特征选取方法性能,将数据集按 3∶1∶1 随机分割为训练集、校验集和独立测试集. 如图 2 所示,首先,根据 Bhattacharyya 距离过滤噪声特征,缩小特征选取范围:在 Colon 训练集里选择 Bhattacharyya 距离较大的前 500 个基因作为特征选取范围;在其他 4 个数据集的训练集里均以 Bhattacharyya 距离较大的前 1000 个基因为特征选取范围. 然后,4 种方法基于相同训练集、相同特征选取范围在 RFE 过程中产生候选特征子集:SVM-RFE 方法基于特征对 SVM 模型的影响对特征排序,生成候选特征子集;BPSO 方法以 SVM 分类器在训练集中 5 倍交叉校验的分类 AUC 值和正确率及特征组合维数倒数为粒子个体适应度函数值;NBPSO 和 ENBPSO 方法通过式(4)计算粒子个体适应度函数值;将 BPSO 方法和 NBPSO 方法选取的特征集合作为各自下一轮特征选取过程的候选特征子集,ENBPSO 方法根据算法 2 生成候选特征基因子集.

4 种方法以 SVM 和 KNN 分类器对相同校验集的分类结果为评价函数从候选特征子集中选择最佳特征集合(即分类特征集合),再以相同的独立测试集测试分类特征集合分类能力. 每种方法重复进行 20 次,得到 20 个分类特征集合,以这些特征集合在独立测试集上分类测试的统计结果比较 4 种特征选取方法的性能.

设置 RFE 过程的特征剔除率为 50%,SVM 核函数为线性核函数,KNN 分类器的 $K=5$,即 5 近邻分类.

PSO 参数是基于经验值基础上,经过多次试验调试确定的. BPSO 方法中 PSO 算法种群规模 $L=100$,$\boldsymbol{k}_{\max}=200$,$\boldsymbol{v}_{\max}=2$,$\boldsymbol{v}_{\min}=-2$,$c_1=c_2=4$,$\omega$ 在进化过程中从 1 逐渐减小至 0.7. NBPSO 方法中 PSO 算法 $\boldsymbol{v}_{\max}=4$,$\boldsymbol{v}_{\min}=-4$,nearMiss = nearHit = 5,其他参数与 BPSO 方法参数相同. ENBPSO 方法设置集成规模 $N=20$,其中 NBPSO 方法中 PSO 算法 $k_{\max}=100$,其他参数与单次 NBPSO 参数相同.

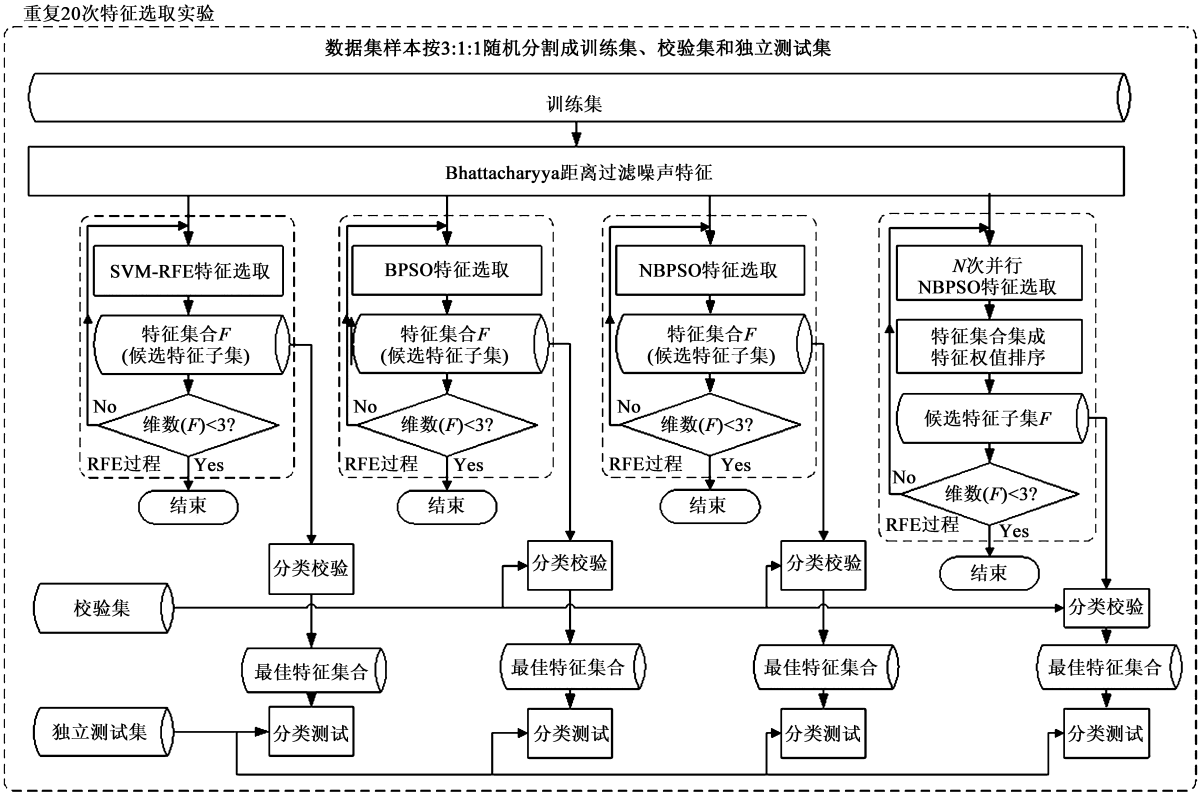


图2　三种方法的特征选取实验流程图

4.2　分类特征集合分类结果

SVM-RFE 方法、BPSO 方法、NBPSO 方法和 ENBPSO 方法在 5 个数据集上选出 20 个分类特征集合,它们在独立测试集里的分类 AUC 值和正确率的平均值和标准差如表 2 所示.

由表 2 知,在 Multiple Myeloma 数据集上,NBPSO 和 ENBPSO 方法所选特征均正确识别所有独立测试集样本类别. 比较 4 种方法所选分类特征集合的分类 AUC 值,在除 Colon 数据集外的其他 4 个数据集上,NBPSO 和 ENBPSO 方法均优于 BPSO 方法和 SVM-RFE 方法;在 DLBCL 和 Prostate 数据集上,ENBPSO 方法优于 NBPSO 方法;在 Acute Leukemia 和 Colon 数据集上,NBPSO 和 ENBPSO 方法选取的分类特征集合的分类 AUC 值互有高低. 比较 4 种方法所选分类特征集合的分类正确率,在 5 个数据集上,NBPSO 和 ENBPSO 方法都优于 BPSO 方法和 SVM-RFE 方法;在 DLBCL、Colon 和 Prostate 数据集上,ENBPSO 方法优于 NBPSO 方法;在 Acute Leukemia 数据集上,NBPSO 和 ENBPSO 方法选取的分类特征集合的分类正确率互有高低.

表 2　分类特征集合在独立测试集上的分类结果统计表

数据集	分类器	分类 AUC 值				分类正确率			
		SVM-RFE	BPSO	NBPSO	ENBPSO	SVM-RFE	BPSO	NBPSO	ENBPSO
DLBCL	SVM	0.907 ±0.092	0.929 ±0.082	0.960 ±0.047	0.963 ±0.043	0.883 ±0.095	0.916 ±0.056	0.947 ±0.041	0.947 ±0.066
	KNN	0.910 ±0.073	0.879 ±0.110	0.917 ±0.095	0.935 ±0.085	0.873 ±0.082	0.833 ±0.102	0.903 ±0.059	0.913 ±0.061
Acute Leukemia	SVM	0.934 ±0.045	0.913 ±0.149	0.959 ±0.045	0.963 ±0.051	0.924 ±0.075	0.878 ±0.116	0.943 ±0.049	0.950 ±0.046
	KNN	0.885 ±0.135	0.857 ±0.161	0.916 ±0.078	0.894 ±0.068	0.858 ±0.111	0.821 ±0.136	0.889 ±0.071	0.864 ±0.103
Multiple Myeloma	SVM	0.980 ±0.049	0.988 ±0.017	1.000 ±0.000	1.000 ±0.000	0.983 ±0.056	0.988 ±0.016	1.000 ±0.000	1.000 ±0.000
	KNN	0.982 ±0.035	0.970 ±0.021	1.000 ±0.000	1.000 ±0.000	0.983 ±0.035	0.988 ±0.017	1.000 ±0.000	1.000 ±0.000
Colon	SVM	0.805 ±0.120	0.809 ±0.125	0.806 ±0.143	0.810 ±0.161	0.735 ±0.120	0.762 ±0.072	0.779 ±0.095	0.804 ±0.081
	KNN	0.832 ±0.095	0.841 ±0.111	0.835 ±0.150	0.829 ±0.162	0.703 ±0.130	0.800 ±0.091	0.808 ±0.086	0.829 ±0.062
Prostate	SVM	0.881 ±0.091	0.910 ±0.086	0.934 ±0.053	0.940 ±0.062	0.824 ±0.083	0.862 ±0.081	0.892 ±0.069	0.920 ±0.063
	KNN	0.883 ±0.089	0.874 ±0.073	0.903 ±0.073	0.933 ±0.072	0.816 ±0.108	0.817 ±0.081	0.835 ±0.082	0.845 ±0.072

5 特征选取方法性能比较

5.1 分类特征集合的分类性能显著性检验

基于单因素非参数方差分析 Kruskal-Walls 检验方法[23],利用 MATLAB 工具箱的 kruskalwallis 函数对 4 种方法所选 20 个分类特征集合在独立测试集上的分类测试结果进行显著性差异分析,以进一步比较四者性能. 每个方法在每个数据集上 80 个独立分类测试数据(20 个 SVM 分类正确率、20 个 SVM 分类 AUC 值、20 个 KNN 分类正确率和 20 个 KNN 分类 AUC 值)组成一组. Kruskal-Walls 检验结果如表 3 所示.

表 3 分类特征集合的分类性能差异性比较表

数据集	SVM-RFE 与 NBPSO	SVM-RFE 与 ENBPSO	BPSO 与 NBPSO	BPSO 与 ENBPSO	NBPSO 与 ENBPSO	显著性水平 α
DLBCL	有	有	有	有	有	2.03%
Acute Leukemia	有	无	有	无	无	4.89%
Multiple Myeloma	有	有	有	有	无	0.01%
Colon	无	无	无	无	无	4.83%
Prostate	有	有	无	有	有	1.35%

综合表 2 和表 3 数据可知,NBPSO 方法和 ENBPSO 方法在 5 个数据集上选出的特征集合比 BPSO 方法和 SVM-RFE 方法选出的特征集合分类能力强;同时,ENBPSO 方法选出的特征集合分类能力比 NBPSO 方法选出的特征集合分类能力略强.

5.2 特征选取方法的时间复杂度

特征选取方法的算法复杂度决定于训练集样本数、特征选取范围、评价函数等因素. 粒子群算法复杂度决定于种群规模、终止条件和适应度函数等因素. 因为算法运行时间与算法复杂度成线性关系,所以相同实验条件下不同算法消耗的时间长短可用于比较算法间的复杂度关系. 表 4 列出 4 种特征选取方法从 5 个数据集选取 20 个分类特征集合所用平均时间.

表 4 特征选取方法运行时间表

数据集	SVM-RFE(s)	BPSO(s)	NBPSO(s)	ENBPSO(s)
DLBCL	2.19	1200	599	7837
AcuteLeukemia	1.89	2404	603	7538
Multiple Myeloma	2.35	3354	879	13682
Colon	0.861	859	400	4503
Prostate	3.91	4208	1020	12783

SVM-RFE 方法用时最短,NBPSO 方法所用时间次之,ENBPSO 方法耗时最长. BPSO 方法消耗的时间主要用于粒子个体适应度函数的 5 倍 SVM 分类交叉校验,NBPSO 方法粒子个体适应度函数基于近邻样本信息计算特征组合的可分性,所用计算量相对 BPSO 较小. ENBPSO 集成方法增加的多次 NBPOS 选取过程提高了特征选取算法的计算量,使之消耗了更多时间. 由上,NBPSO 方法复杂度比 BPSO 方法复杂度低;因为集成方法增加的选取过程提高了 ENBPSO 方法复杂度.

5.3 特征选取方法的稳定性

特征选取方法性能包括分类特征集合分类能力、选取方法稳定性和消耗时间等指标. 所谓稳定性是指特征选取方法在变化的数据环境中所选分类特征集合的相对稳定度.

基于秩相似系数的特征选取方法稳定度指标[15],根据 4 种方法 20 次特征选取实验所得特征集合分析特征选取方法稳定性. 如表 5 所示,SVM-RFE 方法的稳定性低于其他 3 种方法. BPSO 方法在 DLBCL 数据集上的稳定性优于 NBPSO 方法,而 NBPSO 方法在其他 4 个数据集上的稳定性优于 BPSO 方法. ENBPSO 方法在 5 个数据集上的稳定性都优于其他 3 种方法,这也是集成方法用计算复杂度换来稳定性.

表 5 特征选取方法稳定度比较表

数据集	SVM-RFE	BPSO	NBPSO	ENBPSO
DLBCL	0.265	0.466	0.400	0.568
Acute Leukemia	0.305	0.365	0.507	0.612
Multiple Myeloma	0.382	0.315	0.555	0.915
Colon	0.213	0.352	0.433	0.434
Prostate	0.287	0.381	0.450	0.628

随着计算机技术特别是计算机硬件技术发展,科学家们逐渐降低对特征选取方法计算复杂度的要求,越来越关注于特征集合的分类能力和特征选取方法的稳定性. 所以,尽管 ENBPSO 方法在运行时间上比其他 3 种方法消耗多,但鉴于稳定性方面的优势,ENBPSO 方法会有更大应用前景.

6 总结

针对高维数据结构特点,本文提出基于近邻信息的二进制 PSO 特征选取方法. 将基于同类近邻样本和异类近邻样本信息定义的特征组合类别可分性作为粒子个体的适应度函数,用特征组合类别可分性加权的群体历史最佳、粒子历史最佳和粒子邻域内最佳个体信息共同指导粒子运动方向,在特征空间里搜索分类特征集合. 5 个高维数据集上的特征选取实验表明基于近邻样本信息构建的粒子个体适应度函数降低了特征选取算法的复杂度;粒子邻域内最佳个体信息的引入提高了特征选取质量;基于类别可分性的加权集成方法有效提高了 NBPSO 方法的特征选取性能.

NBPSO 方法和 ENBPSO 方法较好地完成了两类样

本特征选取任务,今后将通过式(4)、(5)分析多类别特征组合在类间和类内的相对关系,进一步研究两种方法在高维特征数据集里的多类别特征选取中的应用.

参考文献

[1] Liu H, Sun J, Liu L, et al. Feature selection with dynamic mutual information[J]. Pattern Recognition, 2009, 42(7): 1330 – 1339.

[2] 邹涛,张翠. 概念级误用检测系统的认知能力研究[J]. 电子学报,2004,32(10):1694 – 1697.
Zou Tao, Zhang Cui. A study on apperception ability of concept level misuse detection system[J]. Acta Electronica Sinica, 2004, 32(10): 1694 – 1697. (in Chinese)

[3] 李颖新,刘全金,阮晓钢. 一种肿瘤基因表达数据的知识提取方法[J]. 电子学报,2004,32(9):1479 – 1482.
LI Ying-Xin, LIU Quan-jin, RUAN Xiao-gang. A method for extracting knowledge from tumor gene expression data [J]. Acta Electronica Sinica, 2004, 32(9): 1479 – 1482. (in Chinese)

[4] 边肇祺,张学工. 模式识别[M]. 北京,清华大学出版社,2004. 176 – 203.
Bian Zhaoqi, Zhang Xuegong. Pattern Recognition[M]. Beijing: Tsinghua University Publisher, 2004. 176 – 203. (in Chinese)

[5] Zhang Daoqiang, Chen Songcan, Zhou Zhi-Hua. Constraint score: A new filter method for feature selection with pairwise constraints[J]. Pattern Recognition, 2008, 41(5): 1440 – 1451.

[6] Guyon I, Weston J, Barnhil S, et al. Gene selection for cancer classification using support vector machines[J]. Machine learning, 2002, 46(1 – 3): 389 – 422.

[7] 王树林,王戟,陈火旺,等. 肿瘤信息基因启发式宽度优先搜索算法研究[J]. 计算机学报,2008,31(4):636 – 649.
Wang Shulin, Wang Ji, Chen Huowang, et al. Heuristic breadth-first search algorithm for informative gene selection based on gene expression profiles[J]. Chinese Journal of Computers, 2008, 31(4): 636 – 649. (in Chinese)

[8] Kennedy J, Eberhart R C. Particle swarm optimization[A]. Proceedings of International Conference on Neutral Networks IV[C]. Piscataway NJ: IEEE Service Center, 1995. 1942 – 1948.

[9] 朱大林,詹腾,张屹,等. 多策略差分进化的元胞多目标粒子群算法[J]. 电子学报,2014,42(9):1831 – 1838.
Zhu Da-lin, Zhan Teng, Zhang yi, et al. Cellular multi-objective particle swarm algorithm based on multi-strategy differential evolution[J]. Acta Electronica Sinica, 2014, 42(9): 1831 – 1838. (in Chinese)

[10] Kennedy J, Eberhart RC. A discrete binary version of the particle swarm algorithm[A]. Proceedings of IEEE International Conference on Systems, Man, and Cybernetics [C]. Washington: IEEE, 1997. 4104 – 4109.

[11] Lin SW, Ying KC, Chen SC, et al. Particle swarm optimization for parameter determination and feature selection of support vector machines[J]. Expert Systems with Applications, 2008, 35(4): 1817 – 1824.

[12] Sahua B, Mishra D. A novel feature selection algorithm using particle swarm optimization for cancer microarray data[A]. Proceedings of International Conference on Modelling Optimization and Computing[C]. USA: Procedia Engineering, 2012, 38. 27 – 31.

[13] 李霞,张田文,郭政. 一种基于递归分类树的集成特征基因选取方法[J]. 计算机学报,2004,27(5):675 – 681.
Li Xia, Zhang Tianwen, Guo Zheng. A novel ensemble of feature selection based on recursive partition-tree[J]. ChineseJournal of Computers, 2004, 27(5): 675 – 681. (in Chinese)

[14] Abeel T, Helleputte T, Van de Peer Y, et al. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods[J]. Bioinformatics, 2010, 26(3): 392 – 298.

[15] Saeys Y, Abeel T, Peer YV. Robust feature selection using ensemble feature selection techniques[A]. Proceedings of ECML PKDD' 2008, Part II[C]. LNAI: Springer, 2008, 5212. 313 – 325.

[16] Provost F, Fawcett T. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions[A]. Proceedings of 3rd International Conference on Knowledge Discovery and Data Mining [C]. USA: AAAI Press, 1997. 43 – 48.

[17] Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning[J]. Nat Med, 2002, 8(1): 68 – 74.

[18] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(5439): 531 – 537.

[19] Zhan F, Hardin J, Kordsmeier B, et al. Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells[J]. Blood, 2002, 99(5): 1745 – 1757.

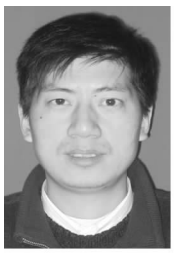
[20] Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays [J]. PNAS USA, 1999, 96(12): 6745 – 6750.

[21] Singh D,Febbo PG,Ross K,et al. Gene expression correlates of clinical prostate cancer behavior[J]. Cancer Cell,2002,1(2):203 – 209.

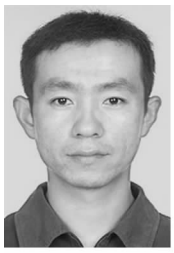
[22] Liu QJ,Zhao ZM,Li YX,et al. Feature selection based on sensitivity analysis of fuzzy ISODATA[J]. Neuro Computing,2012,85(1):29 – 37.

[23] 周品. MATLAB 概率与数理统计[M]. 北京:清华大学出版社,2012. 267 – 270.
Zhou Pin. MATLAB Probability and Mathematical Statistics[M]. Beijing: Tsinghua University Publisher, 2012. 267 – 270. (in Chinese)

作者简介



刘全金 男,1971 年 12 月生于安徽六安,南京航空航天大学理学院博士研究生,安庆师范学院教授,主要研究方向:机器学习、信息处理.
E-mail:liuquanjing666@126.com



李颖新 男,1972 年 9 月生于河北迁安,北京经纬纺机新技术有限公司高级工程师,博士,主要研究方向:机器视觉、机器学习与数据挖掘、生物信息学.
E-mail:linterlee@126.com

赵志敏(通信作者) 女,1955 年 3 月生于辽宁沈阳,南京航空航天大学理学院教授,主要研究方向:现代测量与控制技术、智能计算.
E-mail:nuaazhzhm@126.com

俞晓磊 男,1981 年 10 月生于江苏南京,江苏省标准化研究院高级工程师、南京理工大学博士后,主要研究方向:通信与射频信号处理、电子信号检测技术.
E-mail:nuaaxiaoleiyu@126.com