【文章编号】 2096-2835(2017)02-0214-05 DOI:10.3969/j.issn.2096-2835.2017.02.013

ReliefF 和 APSO 混合降维算法研究

陈俊颖,陆慧娟,严 珂,叶敏超
(中国计量大学 信息工程学院,浙江 杭州 310018)

【摘 要】 降维与分类一直是机器学习的研究热点,在很多领域有着成功的应用.针对基因数据分类存在特征维数过高、冗余数据和高噪声等问题,现提出一种基于 ReliefF 和自适应粒子群(APSO)优化的混合降维算法.即先通过 ReliefF 和 APSO 算法选择特征子集,然后使用超限学习机作为评价函数对基因数据进行分类,最后通过循环迭代得到最优的分类精度.实验证明,混合降维算法与已有的算法相比分类精度更高、更稳定,它适用于基因表达数据降维.

【关键词】 ReliefF 算法;APSO 算法;降维;基因表达数据
【中图分类号】 TP391 【文献标志码】 A

A hybrid dimension reduction algorithm on ReliefF and APSO

CHEN Junying, LU Huijuan, YAN Ke, YE Minchao
(College of Information Engineering, China Jiliang University, Hangzhou 310018, China)

Abstract: Dimensionality reduction and classification are two hot topics in the field of machine learning. We proposed a hybrid feature selection algorithm combining ReliefF and adaptive particle swarm optimization (APSO) for gene data classification, solving the problems of high dimension, redundancy as well as noise. The algorithm extracted the feature subsets by using ReliefF and APSO. The extreme learning machine was used as the evaluation function to classify the gene expression data. The optimized classification accuracy was obtained by recursive substitutions. Experiments show that the proposed hybrid dimensionality reduction algorithm contributes to higher classification accuracy and is more stable than existing algorithms. Consequently, it is an appropriate method for the dimensionality reduction of gene expression data.

Key words: ReleifF algorithm; APSO algorithm; dimensionality reduction; gene expression data

随着生物信息学的发展,发现了一种可以对癌症进行诊断的技术,定义为 DNA 微阵列技术[1-2].在对基因表达数据分类的研究中,发现基因表达数据具有维数高、小样本和非线性等特点[3],而如何获得较高的分类精度、稳定的泛化性能、降低时间复杂度,成为当前基因表达数据分析

【收稿日期】 2017-03-17 《中国计量大学学报》网址:zgjl.cbpt.cnki.net
【基金项目】 国家自然科学基金资助项目(No. 61272315,61602431),浙江省科技厅国际合作专项(No. 2017C34003).
【第一作者简介】 陈俊颖(1992-),男,浙江省杭州人,硕士研究生,主要研究方向为机器学习.E-mail:1820574626@qq.com
通信联系人:陆慧娟,女,教授.E-mail:hjlu@cjlu.edu.cn

的研究重点. 对样本数据的筛选、特征选择、特征提取、分类等都是当今数据挖掘和机器学习中的研究热点. 通过对基因表达数据的挖掘和学习，人们能够清楚的了解差异基因的功能和对其进行干预而引起的结果，并最终将获得的信息作为诊断和治疗的依据. 如今对于基因表达数据分类[4]，主要体现在降维、分类精度以及分类器的稳定性上.

最新的降维方法提出了 MRMD(maximum relevance maximum distance)，即最大相关最大距离. MRMD 选择出和类标具有强相关并且特征之间具有低冗余的特征子集，主要有两部分组成的：一是特征与数据类标之间的相关性，用Pearson 相关系数来计算特征和类标之间的相关性；二是特征之间的冗余性，用三种距离函数(Euclidean 距离、Cosine 距离和 Tanimoto 系数)和的平均值来计算特征之间的冗余性. 而本文的 ReliefF 算法考虑了特征与数据类标之间的相关性.

Relief 算法作为一个高效的特征选择方法，吴艳文[5]等人，分析 Relief 算法及其在聚类应用中存在的问题，提出了一种基于 Relief 算法的特征评价函数，以解决特征权值取值不当对聚类产生的负面影响. 张翔[6]等人为了提高 Relief 特征加权的适应性和鲁棒性，融合间距最大化和极大熵理论，构造了一个结合极大熵原理的间距最大化目标函数，提出了一组鲁棒的 Relief 特征加权算法. 吴红霞[7]等人，提出基于 Relief 和 SVM-RFE 的组合式 SNP 特征选择方法. Relief 虽然应用广泛，但是只能处理两类的数据，因此 1994 年 Kononeill 对其进行了扩展，得到了 ReliefF 算法. ReliefF 算法可以处理多类别问题，能够处理目标属性为连续值的回归问题. 从目前的文献来看，对 ReliefF 算法还有很多值得进一步研究的空间.

粒子群算法，也称粒子群优化算法，是近年来由 J. Kennedy 和 R. C. Eberhart 等[8]研究开发的一种新型进化算法. 在对生物群体观察和研究的基础上，通过群体中每个个体共享群体信息，逐渐完善自身行为轨迹的寻优过程，最终得到最优解. P Ghamisi[9]等人，通过将遗传算法(genetic algorithm，GA)和 PSO 算法进行混合交叉迭代，提出了一种 HGAPSO-Selection 算法. 涂娟娟[10]等人，发现现有的 PSO 算法在迭代的过程中粒子群的多样性会逐渐缺失，提出了一种改进的分期变异 PSO 算法，早期对粒子群进行变异，后期对个体极值和全局极值进行随机扰动，优化神经网络参数及结构，始终保持粒子群的多样性. PSO 特征提取算法由于算法简单、易于实现、参数少、收敛速度快等优点在连续优化问题和离散优化问题中都表现出良好的效果，但其在全局搜索时容易陷入局部最优解而导致早熟收敛、收敛速度慢等问题.

Relief 系列算法通过计算权重来进行特征选择运行速度快，且不限制数据的类型. 粒子群算法能够快速地寻找最优的过程，还能对系统的参数进行优化改进. ReliefF 是特征选择算法，需要丢弃一些基因，这样将会丢失一部分有用的信息，而 PSO 算法是特征提取算法，会浪费资源在冗余和噪声基因上. 然而，APSO 可以提高搜索能力得到更优的解决方案，平衡算法的全局与局部搜索能力，从而提高了算法的多样性与搜索效率[11].

综上所述，本文提出了一种基于 ReliefF 和 APSO 的混合降维算法.

1　ReliefF 算法

ReliefF 算法的核心思想就是特征和数据集类标之间的相关性. ReliefF 算法从训练集中随机选择一个样本 $\boldsymbol{R}$. 先找出与 $\boldsymbol{R}$ 属于同一类且离它最近的样本称为 NearHit，然后再找出与 $\boldsymbol{R}$ 属于不同类且离它最近的样本称为 NearMiss，根据以下算法得到权重：

1)　计算 $\boldsymbol{R}$ 和 NearHit 在第 A 个特征上的距离，S_A^{Hit}.

2)　计算 $\boldsymbol{R}$ 和 NearMiss 在第 A 个特征上的距离，S_A^{Miss}.

3)　比较两个距离 S_A^{Hit} 和 S_A^{Miss}. S_A^{Hit} 大于 S_A^{Miss}，则代表第 A 个特征对区分同类和不同类有用，增加该特征的权重. 而 S_A^{Hit} 小于 S_A^{Miss}，则代表第 A 个特征对区分同类和不同类起负面作用，减小该特征的权重.

4)　重复以上过程 m 次，得到每个特征的平均权重. 其计算公式如下：

$$W(A)=W(A)-\mathrm{diff}(A,\boldsymbol{R},\boldsymbol{H})/m+\mathrm{diff}(A,\boldsymbol{R},\boldsymbol{M})/m. \tag{1}$$

其中 $\mathrm{diff}(A,\boldsymbol{R},\boldsymbol{H})$ 表示样本 $\boldsymbol{R}$ 和它的同类 NearHit 在特征 A 上的差.

重复 m 次以后，每个特征都得到一个平均权

重,平均权重越大,则代表该特征能更好的区分不同类别,而平均权重越小,则代表该特征不能很好的区分不同类别. ReliefF 算法的复杂度随着样本抽样次数 m 和原始特征集个数 N 的增加而线性增加,所以运行的速度快. Relief 系列算法运行速度快,对数据类型没有限制,会给予所有类别相关性高的特征较高的权重,所以属于一种特征权重算法.

2 APSO 算法

粒子群算法,是一种从飞鸟集群捕食行为的研究中发现的. 群体中每个个体共享群体信息,逐渐完善自身行为轨迹的寻优过程,最终得到最优解. 是一种基于迭代的优化算法.

PSO 算法初始化一群随机粒子(随机解). 根据粒子的速度和位置的更新公式,不断进行迭代寻找个体最优解和历史最优解. 在每一次的迭代过程中,每个粒子通过跟踪两个"极值"($p$best, $g$best)来更新自己的位置和速度.

$$\boldsymbol{V}_i^{k+1}=w\boldsymbol{V}_i^k+c_1r_1(\boldsymbol{pb}_i^k-\boldsymbol{X}_i^k)+c_2r_2(\boldsymbol{gb}_i^k-\boldsymbol{X}_i^k), \tag{2}$$

$$\boldsymbol{X}_i^{k+1}=\boldsymbol{X}_i^k+\boldsymbol{V}_i^{k+1}. \tag{3}$$

其中 w 为权值,$\boldsymbol{V}$ 是粒子的速度,$\boldsymbol{X}$ 是粒子的位置,c_1 和 c_2 是学习因子,通常取 $c_1=c_2=2$. 粒子群算法的本质就是群体共享信息来帮助粒子移动到下一个迭代位置,直至找到最优解. 个体充分利用自身信息和以及群体共享的信息来调整自己的位置是粒子群算法的关键所在.

由于 PSO 算法在处理多极值函数问题上经常出现早熟收敛等问题,为增加粒子跳出局部极值的能力,本文提出 APSO 算法在速度更新时加入惯性系数,变化后的公式如下:

$$v_{id}^{k+1}=\omega v_{id}^k+c_1r_1(p\text{best}_{id}^k-x_{id}^k)+c_2r_2(g\text{best}_{id}^k-x_{id}^k), \tag{4}$$

$$\omega=(\omega_{\max}-\omega_{\min})\times\exp(-(\tau\times\frac{k}{K_{\max}})^2)+\omega_{\min}. \tag{5}$$

式(5)中 $\omega_{\max}$,$\omega_{\min}$ 分别为最大、最小惯性系数,$K_{\max}$ 为最大迭代次数,τ 为经验值,一般在[20,55]内取值. 较大的 ω 令 APSO 具有较强的全局搜索能力,ω 较小则更倾向于局部搜索. 惯性系数在初期的时候一般较大,随着算法的迭代次数增加而逐渐减少. 通过改变惯性系数的大小,达到不同的搜索效果. 随着惯性系数的逐渐减小,算法也由初期的全局搜索到后期的局部搜索.

3 基于 ReliefF 和 APSO 的混合降维算法

首先随机产生一个含有 N 个粒子的粒子群,然后使用 ELM 分类器作为评价函数,得到适应度后进行 ReliefF-APSO 混合降维处理,不断迭代最终得到特征子集和最优的分类精度. 本文将其应用于基因数据分类中,通过与其他特征选择方法进行对比,验证该方法的有效性. 图 1 是 ReliefF和 APSO 混合降维算法的框架图.

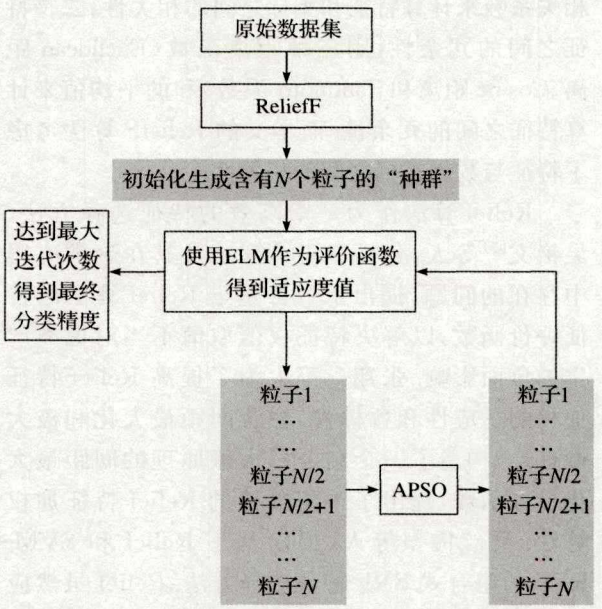

图 1 ReliefF 和 APSO 混合降维算法框架图

Figure 1 Frame chart of the hybrid dimensionality reduction algorithm base on ReliefF and APSO

ReliefF 和 APSO 混合降维算法步骤如下:

Step 1:给定数据集,在训练前,对数据进行 0 均值归一化处理.

Step 2:使用公式(1)进行 ReliefF 算法特征提取,取权值相对较大的前 k 个特征作为 APSO 的训练集和测试集. 实验中 k 取 100.

Step 3:建立基于 APSO 的极限学习机神经网络拓扑结构,设置隐层神经元数目,选择 sigmoid 作为激活函数.

Step 4:产生种群,设置粒子数 N,每个粒子设置为[-1,1]范围内的随机数向量,设置神经元

个数及隐层节点数. 实验中 N 取 20 个.

Step 5：初始化 APSO 的速度与位置变量，设置种群的个体最优位置、群体最优位置.

Step 6：计算每个粒子的适应度值.

Step 7：根据式(4)、(5)更新自适应粒子群的位置和速度.

Step 8：判断是否达到最大迭代次数，若达到，则停止迭代，否则转 step6，继续迭代.

4　实验结果与分析

4.1　数据集

使用的数据集是 UCI 上的基因数据集，如表 1.

表 1　数据集样本数

Table 1　Numble of datasets

数据集	训练集	测试集
Breast	576	192
Colon	62	22
Leukemia	54	18

4.2　预处理(归一化)

因为不同的数据在不同维度上数据的数量级相差过大的话，数值大的数的变化会忽略掉数值小的数的变化. 其次是归一化之后收敛速度快. 在分类算法中，需要使用距离来度量相似性的时候，使用 0 均值标准化会更好. 0 均值归一化是将原始数据集归一化为均值为 0、方差为 1 的数据集，归一化公式如下：

$$Z=\frac{x-\mu}{\sigma^2}. \tag{6}$$

其中 μ,σ^2 分别为原始数据集的均值和方差.

4.3　实验分析

使用 ReliefF 和 APSO 混合降维算法对样本数据进行分类，通过与单一的 APSO、单一的 ReliefF、GA 算法比较分类精度来评价本文降维方法的优劣. ReliefF 和 APSO 混合降维算法简称为 ReliefF-APSO 算法. 在 Breast，Colon 和 Leukemia 3 个基因表达数据集上进行实验，得到的分类精度如表 2，图 2～4.

表 2　不同基因表达数据集下各算法的分类精度

Table 2　Classification accuracy of different gene expression data sets and algorithm

数据集	算法	迭代次数										
		1	5	10	15	20	25	30	35	40	45	50
Breast	APSO	0.631 6	0.684 2	0.736 8	0.736 8	0.842 1	0.842 1	0.842 1	0.842 1	0.842 1	0.842 1	0.842 1
	ReliefF-APSO	0.725 3	0.725 3	0.725 3	0.836 4	0.836 4	0.916 7	0.916 7	0.916 7	0.916 7	0.916 7	0.916 7
	ReliefF	0.421 1	0.684 2	0.578 9	0.421 1	0.578 9	0.368 4	0.473 7	0.526 3	0.315 8	0.315 8	0.526 3
	GA	0.738 1	0.738 1	0.750 0	0.761 9	0.767 9	0.779 8	0.779 8	0.785 7	0.797 6	0.803 6	0.803 6
Colon	APSO	0.545 5	0.636 4	0.636 4	0.636 4	0.727 3	0.727 3	0.818 2	0.818 2	0.818 2	0.818 2	0.818 2
	ReliefF-APSO	0.684 2	0.789 5	0.789 5	0.842 1	0.842 1	0.842 1	0.894 7	0.894 7	0.894 7	0.894 7	0.894 7
	ReliefF	0.750 0	0.545 5	0.583 3	0.666 7	0.833 3	0.727 3	0.916 7	0.833 3	0.666 7	0.916 7	0.500 0
	GA	0.678 3	0.678 3	0.693 3	0.693 3	0.696 7	0.696 7	0.704 2	0.718 3	0.718 3	0.723 5	0.735 9
Leukemia	APSO	0.631 6	0.631 6	0.631 6	0.631 6	0.657 9	0.657 9	0.657 9	0.657 9	0.684 2	0.684 2	0.684 2
	ReliefF-APSO	0.631 6	0.736 8	0.842 1	0.868 4	0.894 7	0.894 7	0.947 3	0.947 3	0.973 7	1.000 0	1.000 0
	ReliefF	0.473 7	0.605 3	0.631 6	0.578 9	0.684 2	0.631 6	0.657 9	0.447 4	0.605 3	0.552 6	0.578 9
	GA	0.614 1	0.614 1	0.666 7	0.684 2	0.701 8	0.754 4	0.771 9	0.789 5	0.786 5	0.796 7	0.803 5

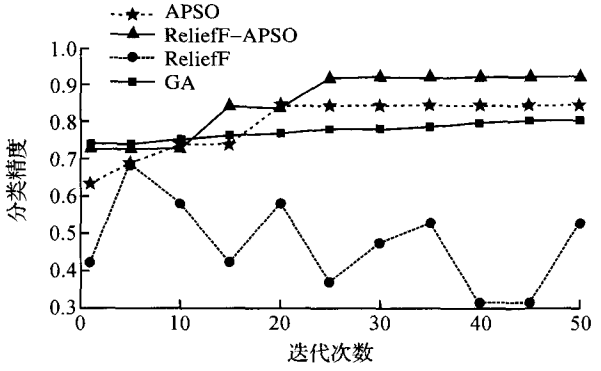

图 2　Breast 数据集下各算法分类精度

Figure 2　Classification accuracy on Breast dataset

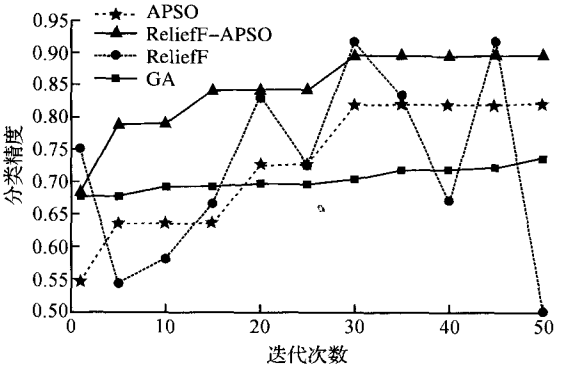

图 3　Colon 数据集下各算法分类精度

Figure 3　Classification accuracy on Colon dataset

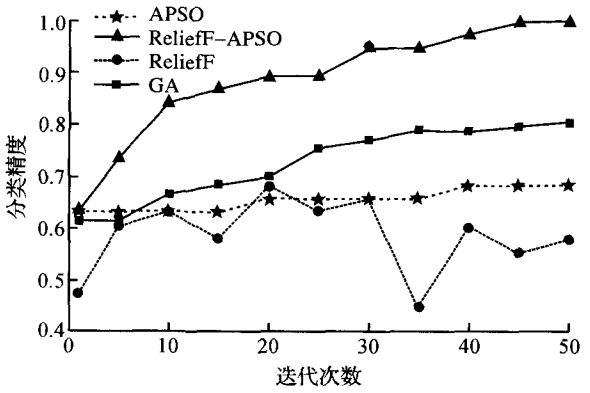

图 4 Leukemia 数据集下各算法分类精度

Figure 4 Classification accuracy on Leukemia dataset

由图 2、图 3、图 4 可知，在基因数据集上，单一的 ReliefF 算法随着迭代次数增加，分类精度一直处于上下波动的状态，极不稳定. 而用 ReliefF-APSO、APSO、GA 算法使用 ELM 算法进行分类，算法的分类精度不再上下波动. 并且随着迭代次数增加，算法的分类精度稳步上升. 本文提出的 ReliefF-APSO 算法在 Breast、Colon、Leukemia数据集上都体现了良好的分类精度，且具有良好的稳定性. 这说明 ReliefF-APSO 算法是一种有效的降维算法.

5 结 语

采用 ReliefF-APSO 算法，先通过ReliefF算法进行特征选择，选取含有信息更多的特征，然后利用 APSO 再进行特征提取. 本算法能够快速的缩小全局搜索范围，具有平衡了算法的全局与局部搜索能力，同时提高了算法的多样性与搜索效率，减小了算法陷入局部最优的可能. 通过与已有类似算法在不同数据集上进行分类对比实验，表明本文提出的方法具有更好的分类性能，在基因数据降维等领域有较好的应用前景，可以做进一步深入研究.

【参 考 文 献】

[1] BRAZMA A, VILO J. Gene expression data analysis [J]. **Microbes & Infection**, 2000, 480(1): 17-24.

[2] SHERLOCK G. Analysis of large-scale gene expression data [J]. **Current Opinion in Immunology**, 2000, 12(2): 201-205.

[3] HELLER M J. DNA Microarray Technology: Devices, Systems, and Applications [J]. **Annual Review of Biomedical Engineering**, 2002, 4(1): 129-153.

[4] 陆慧娟. 基于基因表达数据的肿瘤分类算法研究[D]. 徐州：中国矿业大学，2012.
LU H J. **A Study of Tumor Classification Algorithms Using Gene Expression Data**[D]. Xuzhou: China University of Mining and Technology, 2012

[5] 吴艳文，胡学钢，陈效军. 基于 Relief 算法的特征学习聚类[J]. 合肥学院学报(自然科学版)，2008，18(2)：45-48.
WU Y W, Hu X, CHEN X J. Feature learning clustering based on Relief algorithm[J]. **Journal of Hefei University (Natural Science Edition)**, 2008, 18(2): 45-48.

[6] 张翔，邓赵红，王士同，等. 极大熵 Relief 特征加权[J]. 计算机研究与发展，2011，48(6)：1038-1048.
ZHANG X, DENG Z H, WANG S T, et al. Maximum entropy Relief feature weighting[J]. **Journal of Computer Research and Development**, 2011, 48(6): 1038-1048.

[7] 吴红霞，吴悦，刘宗田，等. 基于 Relief 和 SVM-RFE 的组合式 SNP 特征选择[J]. 计算机应用研究，2012，29(6)：2074-2077.
WU H X, WU Y, LIU Z T, et al. Combined SNP feature selection based on relief and SVM-RFE[J]. **Application Research of Computers**, 2012, 29(6): 2074-2077.

[8] KENNEDY J, EBERHART R. Particle swarm optimization [C]//**Proceedings of IEEE International Conference on Neural Networks**. Piscataway: IEEE, 1995: 1942-1948.

[9] GHAMISI P, BENEDIKTSSON J A. Feature selection based on hybridization of genetic algorithm and particle swarm optimization [J]. **IEEE Geoscience & Remote Sensing Letters**, 2015, 12(2): 309-313.

[10] 涂娟娟. PSO 优化神经网络算法的研究及其应用[D]. 镇江：江苏大学，2013.
TU J J. **Research on Learning Algorithm of Neural Network Optimized with PSO and its Application**[D]. Zhenjiang: Jiangsu University, 2013.

[11] 陈晓青，陆慧娟，郑文斌，等. 自适应混沌粒子群算法对极限学习机参数的优化[J]. 计算机应用，2016，36(11)：3123-3126.
CHEN X Q, LU H J, ZHENG W B, et al. Optimization of extreme learning machine parameters by adaptive chaotic particle swarm optimization algorithm[J]. **Journal of Computer Applications**, 2016, 36(11): 3123-3126.