基于随机森林的特征选择算法

姚登举[1,2],杨 静[1],詹晓娟[3]

(1. 哈尔滨工程大学 计算机科学与技术学院,哈尔滨 150001;2. 哈尔滨理工大学 软件学院,哈尔滨 150040;
3. 黑龙江工程学院 计算机科学与技术学院,哈尔滨 150050)

摘 要:提出了一种基于随机森林的封装式特征选择算法RFFS,以随机森林算法为基本工具,以分类精度作为准则函数,采用序列后向选择和广义序列后向选择方法进行特征选择。在UCI数据集上的对比实验结果表明,RFFS算法在分类性能和特征子集选择两方面具有较好的性能。

关键词:人工智能;随机森林;特征选择;封装式

中图分类号:TP18 **文献标志码**:A **文章编号**:1671-5497(2014)01-0137-05

DOI:10.13229/j.cnki.jdxbgxb201401024

Feature selection algorithm based on random forest

YAO Deng-ju[1,2], YANG Jing[1], ZHAN Xiao-juan[3]

(1. *College of Computer Science and Technology, Harbin Engineering University, Harbin* 150001, *China*; 2. *School of Software, Harbin University of Science and Technology, Harbin* 150040, *China*; 3. *College of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin* 150050, *China*)

Abstract: A feature selection algorithm based on random forest (RFFS) is proposed. This algorithm adopts random forest algorithm as the basic tool, the classification accuracy as the criterion function. The sequential backward selection and generalized sequential backward selection methods are employed for feature selection. The experimental results on UCI datasets show that the RFFS algorithm has better performance in classification accuracy and feature selection subset than the other methods in literatures.

Key words: artificial intelligence; random forest; feature selection; wrapper

0 引 言

图像处理、信息检索以及生物信息学等技术的发展,产生了以超大规模特征为特点的高维数据集。如何有效地从高维数据中提取或选择出有用的特征信息或规律,并将其分类识别已成为当今信息科学与技术所面临的基本问题[1]。特征选择是指从原始特征集中选择使某种评估标准最优的特征子集,以使在该最优特征子集上所构建的分类或回归模型达到与特征选择前近似甚至更好

收稿日期:2012-08-21.

基金项目:国家自然科学基金项目(61073043,61073041);黑龙江省自然科学基金项目(F200901,F201313);哈尔滨市科技创新人才研究专项项目(2011RFXXG015,2010RFXXG002,2013RFQXJ114);高等学校博士学科点专项科研基金项目(20112304110011).

作者简介:姚登举(1980-),男,博士研究生,讲师.研究方向:人工智能,数据挖掘,模式识别.
E-mail:ydkvictory@163.com

的预测精度。Davies 证明寻找满足要求的最小特征子集是 NP 完全问题[2]。在实际应用中,通常是通过采用启发式搜索算法,在运算效率和特征子集质量间找到一个好的平衡点,即近似最优解。

随机森林(Random forest,RF)[3]是一种集成机器学习方法,它利用随机重采样技术 bootstrap 和节点随机分裂技术构建多棵决策树,通过投票得到最终分类结果。RF 具有分析复杂相互作用分类特征的能力,对于噪声数据和存在缺失值的数据具有很好的鲁棒性,并且具有较快的学习速度,其变量重要性度量可以作为高维数据的特征选择工具,近年来已经被广泛应用于各种分类、预测、特征选择以及异常点检测问题中[4-7]。

特征选择算法根据所采用的特征评价策略可以分为 Filter 和 Wrapper 两大类[8]。Filter 方法独立于后续采取的机器学习算法,可以较快地排除一部分非关键性的噪声特征,缩小优化特征子集搜索范围,但它并不能保证选择出一个规模较小的优化特征子集。Wrapper 方法在筛选特征的过程中直接用所选特征子集来训练分类器,根据分类器在测试集的性能表现来评价该特征子集的优劣,该方法在计算效率上不如 Filter 方法,但其所选的优化特征子集的规模相对要小一些。

本文以随机森林算法为基本工具研究 Wrapper 特征选择方法,利用随机森林分类器的分类准确率作为特征可分性判据,基于随机森林算法本身的变量重要性度量进行特征重要性排序,利用序列后向选择方法(Sequential backward selection,SBS)和广义序列后向选择方法(Generalized sequential backward selection,GSBS)选取特征子集。实验结果表明,相比于文献中[9-10]已有的特征选择算法,本文的算法在性能上有较大的提高。

1 随机森林

定义 1 随机森林[3]是一个由一组决策树分类器 $\{h(X,\theta_k),k=1,2,\cdots,K\}$ 组成的集成分类器,其中 $\{\theta_k\}$ 是服从独立同分布的随机向量,K 表示随机森林中决策树的个数,在给定自变量 X 下,每个决策树分类器通过投票来决定最优的分类结果。

随机森林是许多决策树集成在一起的分类器,如果把决策树看成分类任务中的一个专家,随机森林就是许多专家在一起对某种任务进行分类。

生成随机森林的步骤如下:

(1)从原始训练数据集中,应用 bootstrap 方法有放回地随机抽取 K 个新的自助样本集,并由此构建 K 棵分类回归树,每次未被抽到的样本组成了 K 个袋外数据(Out-of-bag,OOB)。

(2)设有 n 个特征,则在每一棵树的每个节点处随机抽取 m_{try} 个特征($m_{\text{try}}\leqslant n$),通过计算每个特征蕴含的信息量,在 m_{try} 个特征中选择一个最具有分类能力的特征进行节点分裂。

(3)每棵树最大限度地生长,不做任何剪裁。

(4)将生成的多棵树组成随机森林,用随机森林对新的数据进行分类,分类结果按树分类器的投票多少而定。

定义 2 给定一组分类器 $h_1(X),h_2(X),\cdots,h_k(X)$,每个分类器的训练集都是从原始的服从随机分布的数据集 (Y,X) 中随机抽样所得,余量函数(Margin function)定义为

$$mg(X,Y)=av_kI(h_k(X)=Y)-\max_{j\neq Y}av_kI(h_k(X)=j) \tag{1}$$

式中:$I(\cdot)$ 是示性函数。

余量函数用于度量平均正确分类数超过平均错误分类数的程度,余量值越大,分类预测越可靠。

定义 3 泛化误差定义为

$$PE^*=P_{X,Y}(mg(X,Y)<0) \tag{2}$$

式中:下标 X、Y 表示概率 P 覆盖 X、Y 空间。

在随机森林中,当决策树分类器足够多,$h_k(X)=h(X,\theta_k)$ 服从强大数定律。

定理 1 随着随机森林中决策树数量的增加,所有序列 $\theta_1,\theta_2,\cdots,\theta_k,PE*$ 几乎处处收敛于

$$P_{X,Y}\{P_\theta(h(X,\theta)=Y)-\max_{j\neq Y}P_\theta(h(X,\theta)=j)<0\} \tag{3}$$

定理 1 表明随机森林不会随着决策树的增加而产生过拟合问题,但可能会产生一定限度内的泛化误差。

变量重要性评估是随机森林算法的一个重要特点。随机森林程序通常提供 4 种变量重要性度量。本文采用基于袋外数据分类准确率的变量重要性度量。

定义 4　基于袋外数据分类准确率的变量重要性度量[7]定义为袋外数据自变量值发生轻微扰动后的分类正确率与扰动前分类正确率的平均减少量。

假设有 bootstrap 样本 $b=1,2,\cdots,B$，B 表示训练样本个数，特征 X_j 的基于分类准确率的变量重要性度量 $\overline{D_j}$ 按照下面的步骤计算：

(1)设置 $b=1$，在训练样本上创建决策树 T_b，并将袋外数据标记为 L_b^{oob}。

(2)在袋外数据上使用 T_b 对 L_b^{oob} 数据进行分类，统计正确分类的个数，记为 R_b^{oob}。

(3)对于特征 X_j，$j=1,2,\cdots,N$，对 L_b^{oob} 中的特征 X_j 的值进行扰动，扰动后的数据集记为 L_{bj}^{oob}，使用 T_b 对 L_{bj}^{oob} 数据进行分类，统计正确分类的个数，记为 R_{bj}^{oob}。

(4)对于 $b=2,3,\cdots,B$，重复步骤(1)～(3)。

(5)特征 X_j 的变量重要性度量 $\overline{D_j}$ 通过下面的公式进行计算：

$$\overline{D_j}=\frac{1}{B}\sum_{i=1}^{B}(R_b^{oob}-R_{bj}^{oob}) \tag{4}$$

定义 5　随机森林算法的分类准确率定义为

$$Accuracy=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{5}$$

式中：TP(true positive)代表正确的肯定；TN(true negative)代表正确的否定；FP(false positive)代表错误的肯定；FN(false negative)代表错误的否定。

2　本文算法 RFFS

2.1　算法描述

本文提出了一种基于随机森林的 Wrapper 特征选择方法 RFFS，利用随机森林算法的变量重要性度量对特征进行排序，然后采用序列后向搜索方法，每次从特征集合中去掉一个最不重要(重要性得分最小)的特征，逐次进行迭代，并计算分类正确率，最终得到变量个数最少、分类正确率最高的特征集合作为特征选择结果。为了保证实验结果的稳定性，本文采用了 10 折交叉验证方法，在每一次迭代中，将数据集划分成 10 等份，利用其中的 9 份作为训练集用于构建随机森林分类器，剩余的 1 份作为验证集数据进行验证。在 10 折交叉验证过程中，选择测试集上分类准确率最高的一次迭代产生的变量重要性排序作为删除特征的依据，将 10 次迭代的平均分类准确率作为该轮迭代的分类精度。具体过程如算法 1 所示。

算法 1　基于随机森林的特征选择算法 RFFS

输入：原始数据集 S

输出：验证集上的最大分类正确率 TGMaxAcc 及其对应的特征集合 FGSort

步骤：

1. 初始化
1.1 读入原始数据集 S
1.2 设置 TGMaxAcc=0
2. For(ft in N−2)
2.1　将数据集 S 随机划分成 10 等份
2.2　设置局部最大分类准确率 TLMaxAcc=0
2.3　设置局部平均分类准确率 TLMeanAcc=0
2.4　初始化 10 折交叉验证中每次迭代的分类准确率
　　TLAcc[1：10]=0
2.5　For(i in 1：10)
2.5.1　在 S 上运行 randomForest 创建分类器
2.5.2　在测试集上执行 predict 进行分类
2.5.3　比较分类结果与观测值，计算 TLAcc
2.5.4　计算 TLMeanAcc = TLMeanAcc + TLAcc[i]/10
2.5.5　If(TLMaxAcc<=TLAcc[i])
2.5.6　　则 TLMaxAcc=TLAcc[i]
2.5.7　　对特征按变量重要性排序并存为 FSort
2.6　If(TGMaxAcc<=TLMeanAcc)
　　则 TGMaxAcc=TLMeanAcc
　FGSort=FSort
2.7　从 FSort 中去掉重要性得分最低的一个特征，得到新的数据集 S
3. 输出结果
3.1 输出全局最高分类准确率 TGMaxAcc
3.2 输出全局最高分类准确率对应的特征集合 FGSort

注：ft 代表循环变量，N 代表数据集中所有特征个数。

2.2　时间复杂度分析

本文所提出的随机森林特征选择方法中基分类器选择 CART 算法。假设训练数据集的特征维数为 m，训练样本个数为 n，CART 算法的时间复杂度为 $O(mn(\log n)^2)$。随机森林在构建 CART 树的过程中，从 m 个特征中随机选择 m_{try} 个特征计算信息增益，并且对树的生长不进行剪枝，故训练每一个基分类器的计算时间小于 $O(mn(\log n)^2)$，假设随机森林中基分类器的个数为 k 个，则随机森林算法的时间复杂度可以近似为 $O(kmn(\log n)^2)$。在本实验中，采用序列后

向选择策略进行特征选择需要循环 $m-2$ 次,每一轮循环中采用 10 折交叉验证,需运行随机森林算法 10 次,每轮循环需对特征子集进行排序,采用快速排序算法的平均时间复杂度为 $O(m\log m)$,根据排序后的特征集合生成新的训练数据集需要进行 $m-2$ 次,每次计算时间为常数,故本算法总的时间复杂度可以近似表示为

$$O((m-2)*(10*O(kmn(\log n)^2)+O(m\log m)+m-2))\approx O(km^2n(\log n)^2) \tag{6}$$

由式(6)可见,RFFS 算法的时间复杂度与特征维数 m 成近似平方关系,与数据集样本个数 n 成近似立方关系,对于高维小样本数据,运算时间是可以接受的,算法具有较好的扩展性。

3 实验结果及分析

3.1 实验数据与方法

为便于比较,从 UCI 数据集中选取了 wdbc、breast-cancer-wisconsin、pima-indians-diabetes 和 heart disease4 个数据集进行测试。表 1 列出了这些数据集的特征,数据集维数从几个到数十个不等。

表 1 取自 UCI 的数据集

Table 1 Dataset from UCI

序号	数据集	维数	样本个数	分类数目
1	WDBC	31	569	2
2	Breast	9	699	2
3	Diabetes	8	768	2
4	Heart	13	270	2

本文算法采用 R 语言进行实现,随机森林核心算法采用 R 软件中的 randomForest 程序包,其中 m_{try} 参数取 Breiman 建议的默认值 $\sqrt{n}$(n 为训练数据集中特征的个数),n_{tree} 参数设置为 1000。实验的硬件环境为 Intel Core(TM)2 Duo CPU E4600@2.40 GHz,3.75 G 的内存,操作系统为 Microsoft Windows 7,绘图软件采用 Matlab7.0。

3.2 实验结果分析

本文算法 RFFS 在搜索实现最大分类准确率的特征子集时采用的是序列后向搜索策略,特征选择过程和结果如图 1 所示。从图 1 可以看出,随着不重要特征(在随机森林变量重要性排序中排在最后的特征)的依次删除,分类准确率整体上呈现逐步提高的趋势,这主要是因为不相关特征和冗余特征的消除提高了分类器性能;当分类准确率到达最高值 97.98% 后又开始呈现下降趋势,则是因为有用的特征被消除,降低了分类器的性能。这说明了本文算法能够有效地识别并消除冗余特征和不相关特征,从而提高分类器的分类性能。

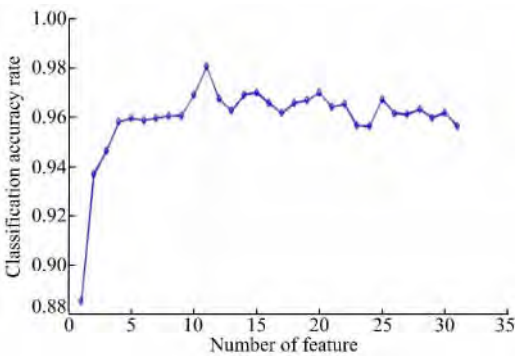

图 1 分类精度与特征个数之间的关系

Fig. 1 Relationship between classification accuracy and feature number

表 2 列出了 RFFS 算法和 CBFS 算法、AMGA 算法在不同实验数据集上的性能比较,其中 SF 列表示选出的最优特征子集中的特征个数,Acc 列表示算法在实验数据集上的分类性能。CBFS 算法、AMGA 算法在相应数据集上的实验数据分别来自文献[9]和文献[10],"—"表示该算法在相应数据集上没有进行实验。

表 2 不同算法的性能比较

Table 2 Performance comparison of different algorithms

数据集	CBFS[9]		AMGA[10]		RFFS	
	SF	Acc	SF	Acc	SF	Acc
WDBC	—	—	—	—	11	0.980
Breast	6	0.943	6	0.991	6	0.982
Diabetes	4	0.670	5	0.804	5	0.811
Heart	9	0.791	9	0.915	6	0.923

从表 2 可以看出,RFFS 算法在 Breast、Diabetes、和 Heart 数据集上的分类正确率分别为 98.2%、81.1%、92.3%,选择出的特征个数分别为 6、5、6。与 CBFS 算法相比,RFFS 算法在特征个数基本相等或者更少的情况下,分类性能明显优于 CBFS。RFFS 算法在 Breast 数据集上选择的特征个数与 AMGA 算法相等,分类性能略低于 AMGA 算法;在 Diabetes 数据集上,RFFS 算法选择的特征个数与 AMGA 算法相等,分类性能略高于 AMGA 算法;在 Heart 数据集上,

RFFS 算法比 AMGA 算法选择了更少的特征数目，却获得了更高的分类精度。从整体上看，本文方法优于文献[9]和文献[10]中的方法。实验结果表明，RFFS 算法不仅能够选择出较优的特征子集，而且能够获得较高的分类性能。

另外，本文算法也可以容易地扩展为使用广义后向搜索策略进行最优子集搜索，为了获得最好的分类效果，本文对删除“最不重要特征”时采用的不同步长进行了实验，结果如表 3 所示。从表 3 可以看出，在所有 6 组实验中（依次删除 1 个到 6 个特征），最高的分类性能是在每次删除一个特征时获得。需要说明的是，每次删除一个特征并不是在所有数据集上的最优选择，由于本文所涉及的数据集特征数目还不是特别高，所以每次删除一个特征有助于获取最优子集。当数据集特征数目非常高时，每次删除一个特征就不再适用，因为大量的特征数目将会大大增加时间开销，并且不能快速有效地消除冗余和不相关特征。如何对采用广义后向搜索时的 L 值进行设置，将是下一步研究的方向。

表 3　每次删除不同个数特征的实验结果

Table 3　Experiment result when deleting different number of features in each time

L	选择出的最优特征	分类正确率
1	V30,V16,V26,V24,V23,V25,V10,V29,V4,V6,V27	0.9798
2	V24,V30,V16,V26,V10,V25,V23	0.9559
3	V30,V26,V16,V10,V24,V23,V25,V4,V29,V6	0.9650
4	V30,V16,V26,V24,V23,V25,V4,V10,V29,V6,V27	0.9671
5	V30,V26,V16,V25,V23,V10	0.9450
6	V30,V24,V16,V26,V10,V25,V23	0.9639

4　结束语

提出了一种基于随机森林的封装式特征选择算法，该算法利用随机森林算法的变量重要性度量对特征进行排序，采用后向序列搜索方法寻找能够训练最优性能分类器的特征子集。实验结果表明本文的特征选择算法可以获得较好的分类性能和特征子集，与以前文献中的方法[9-10]相比具有一定的优势。如何在高维数据集中确定广义后向搜索方法中的 L 值，是下一步的研究内容。

参考文献：

[1] 蒋胜利. 高维数据的特征选择与特征提取研究[D]. 西安：西安电子科技大学计算机学院，2011.
Jiang Sheng-li. Research on feature selection and feature extraction for high-dimensional data[D]. Xi'an: School of Computer Science and Engineering, Xidian University, 2011.

[2] Davies S, Russl S. NP-completeness of searches for smallest possible feature sets[C]// Proceedings of the AAAI Fall Symposiums on Relevance, Menlo Park, 1994: 37-39.

[3] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[4] Strobl Carolin, Boulesteix Anne-Laure, Kneib Thomas, et al. Conditional variable importance for random forests[J]. BMC Bioinformatics, 2008, 9(1): 1-11.

[5] Reif David M, Motsinger Alison A, McKinney Brett A, et al. Feature selection using a random forests classifier for the integrated analysis of multiple data types[C]// IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology, 2006: 171-178.

[6] Mohammed Khalilia, Sounak Chakraborty, Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest[J]. BMC Medical Informatics and Decision Making, 2011, 11(7): 51-58.

[7] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: a survey and results of new tests[J]. Pattern Recognition, 2011, 44(2): 330-349.

[8] Inza I, Larranaga P, Blanco R. Filter versus wrapper gene selection approaches in DNA microarray domains[J]. Artificial Intelligence in Medicine, 2004, 31(2): 91-103.

[9] 蒋盛益，郑琪，张倩生. 基于聚类的特征选择方法[J]. 电子学报，2008，36(12)：157-160.
Jiang Sheng-yi, Zheng Qi, Zhang Qian-sheng. Clustering-based feature selection[J]. Acta Electronica Sinica, 2008, 36(12): 157-160.

[10] 刘元宁，王刚，朱晓冬，等. 基于自适应多种群遗传算法的特征选择[J]. 吉林大学学报：工学版，2011，41(6)：1690-1693.
Liu Yuan-ning, Wang Gang, Zhu Xiao-dong, et al. Feature selection based on adaptive multi-population genetic algorithm[J]. Journal of Jilin University(Engineering and Technology Edition), 2011, 41(6): 1690-1693.