第38卷 第4期 华 中 科 技 大 学 学 报(自然科学版) Vol. 38 No. 4
2010年 4月 J. Huazhong Univ. of Sci. & Tech. (Natural Science Edition) Apr. 2010

基于随机森林相似度矩阵差异性的特征选择

周绮凤 洪文财 杨 帆 罗林开

(厦门大学 自动化系, 福建 厦门 361005)

摘要: 将随机森林的相似度矩阵看做一种特殊的核度量,利用该度量对模型参数的鲁棒性和特征变化的敏感性,提出一种特征选择的方法. 采用相似度矩阵,计算训练样本类内和类间相似性比率. 再利用特征值随机置换技术,将相似性比率的变化量作为特征重要性度量指标,从而对所有特征进行排序. 试验结果表明,该方法能充分利用全部样本的信息,有效地进行特征选择,且其性能优于基于袋外数据误差率估计的特征选择方法.

关 键 词: 特征选择; 度量; 差异性; 相似度矩阵; 随机森林; 随机置换

中图分类号: TP391 **文献标识码**: A **文章编号**: 1671-4512(2010)04-0058-04

Feature selection of random forest-based proximity matrix difference

Zhou Qifeng Hong Wencai Yang Fan Luo Linkai

(Department of Automation, Xiamen University, Xiamen 361005, Fujian China)

Abstract: A feature selection method is proposed, after analyzing proximity matrix's to random forest model and its sensitiveness to the variation of features. Proximity matrix is taken as a special kernel measurement to compute the proximity ratio between inner-class and the inter-class, then permutes the values of feature randomly and the difference of proximity ratio was takes as the assessment criterion for feature importance. The process yields a ranking for all features. Experimental results show that the method achieves good effects and performs better than that of the method based on out-of-bag (OOB) error rate.

Key words: feature extraction; measurements; differentiation; proximity matrix; random forest; random permutation

特征选择是模式识别核心问题之一. 特征选择的过程可以视为一个启发式的搜索问题,即在全空间搜索一个具体的特征子集[1]. 然而,最优特征子集选择是一个 NP 问题[2]. 因此,寻找一个较好的近似算法具有现实意义. 目前,实际应用中出现了许多基于智能学习的特征选择算法[3]. 如广泛应用于基因选择等问题中的基于支持向量机的递归特征消除[4](SVM-RFE)算法. 该方法评价每个特征对支持向量机分类性能的影响,递归地消除不重要的特征,最终得到一个较优的特征子集. 此外,采用随机森林(RF)的变量重要性进行特征选择也逐渐成为研究热点[5]. 随机森林是 Breiman 于 2001 年提出的一个新的组合分类器[6~9]. 它首先采用 Bagging 方法制造有差异的训练样本集,并以分类回归树作为元分类器,当构建单棵树时,采用类似随机子空间划分的策略,随机地选择特征对内部节点进行属性分裂. 这种"双随机"的策略在各子分类器之间形成较大的差异性,使得随机森林具有优越的分类性能,成为最成功的集成学习方法之一.

SVM-RFE 对非线性可分及多分类问题进行特征选择存在很大的局限性. 已有的 RF 算法进行特征选择时,对 OOB 数据的准确率变化有时并不敏感. 据此,本文对随机森林在建模过程中产生的相似度矩阵进行了深入分析,提出一种基于相似度矩阵差异性的特征选择方法. 并通过理论

收稿日期: 2009-08-14.
作者简介: 周绮凤(1976-),女,讲师, E-mail: zhouqf@xmu.edu.cn.
基金项目: 福建省自然科学基金资助项目 (2009J05153).

分析与实验比较,验证了该方法能有效地对特征进行排序,从而选出合适的特征子集.

1　基于相似度矩阵差异性的特征选择

1.1　随机森林的样本相似性度量

随机森林在建模的同时,还提供了样本相似性度量,即相似度矩阵(简记为 Prox 矩阵). 当用一棵树对所有数据进行判别时,这些数据最终都将达到该树的某个叶节点上. 可以用两个样本在每棵树的同一个节点上出现的频率大小,来衡量这两个样本之间的相似程度,或两个样本属于同一类的概率大小. Prox 矩阵生成过程如下:**a**. 对于样本数为 n 的训练集,首先生成一个 $n \times n$ 的零元素矩阵 Prox,记为 $P = \{p_{ij}\}$;**b**. 对于任意两个样本 x_i 和 x_j,若它们出现在所建树的同一个叶节点上,则 $p'_{ij} = p_{ij} + 1$;**c**. 重复上述过程直至 m 棵树全部建好,得到相应的矩阵;**d**. 进行归一化处理 $p'_{ij} = p_{ij}/m$ $(i, j = 1, 2, \cdots, n)$,得到最后的 Prox 矩阵.

由上述过程可以看出 Prox 矩阵是一个主对角线为 1 的实对称方阵. 若数据集中某一类的样本数较多,则该类中的样本所对应的行通常都包含较多接近 1 的元素,而那些包含较多接近零元素的行所对应的样本和其他样本的相似度较小. 所以,Prox 矩阵是一种合适的样本间相似关系度量. 随机森林在计算 Prox 矩阵的过程中,相当于一种特征映射工具,它将原始空间的样本映射到相似性空间中. 因此,也可以将 RF 相似性度量看成是一种特殊的核度量,且这种离散化的取值方式使得样本在相似度空间的差异间隔变大,样本能被更有效地区分开来.

1.2　Prox 矩阵性质研究

a. Prox 矩阵对 RF 的鲁棒性. 随机森林在建模过程中主要涉及到的参数是树的个数 T 及在每个内部节点随机抽取的候选特征个数 m_{try}. m_{try} 一般取值为 $[\sqrt{l}]$, l 是特征的个数,只要在训练过程中特征数保持不变,就可以认为 m_{try} 是一个常量. 树的个数一般取一个适中的值如 1 000,2 000 等. 而随机森林在训练过程中采用 Bagging 的方法制造有差异的训练样本集,当构建单棵树时,随机地选择特征对内部节点进行属性分裂. 即每次建树的过程中,训练样本和分裂属性都是随机选择的. 为此,本文设计如下的统计量,衡量这种"双随机性"对每次建模后 Prox 矩阵的影响.

设任意两个 $n \times n$ 相似度矩阵 P_1 和 P_2 的差异为

$$V_2 = P_1 - P_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} | p_{1i,j} - p_{2i,j} |, \quad (1)$$

式中 $p_{i,j}$ 表示相似度矩阵 P 的第 i 行第 j 列元素. 定义 m 个 Prox 矩阵的平均差异为

$$V_m = \frac{1}{m-1}\sum_{i=1}^{m}(P_i - \overline{P}), \quad (2)$$

式中 $\overline{P}$ 是 m 个 Prox 矩阵的均值, $\overline{P} = (1/m) \cdot \sum_{i=1}^{m} P_i$.

利用统计量 d 在 iris,zoo,SAheart,dermatology 4 个 UCI 标准数据库上,对全部样本生成的 Prox 矩阵的鲁棒性进行试验,试验结果如图 1 所示.

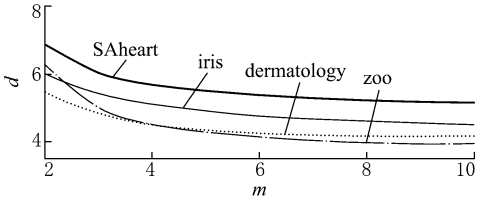

图 1　Prox 矩阵鲁棒性分析

图 1 显示,在 4 个数据库上统计量 d 都非常小,说明对同一数据集,RF 虽然是随机选择训练样本和分裂属性,但其产生的 Prox 矩阵具有统计不变性,因此 RF 提供的相似性度量是鲁棒的.

b. Prox 矩阵对特征的敏感性. 为了比较 OOB 误差率和 Prox 矩阵对特征改变的敏感性,本文以 dermatology 数据集中按特征重要性排列的前 5 个特征(21,22,28,5,12)和后 5 个特征(20,3,34,7,30,31)分别随机重排,观察 OOB 误差率和 Prox 矩阵的相对变化.

图 2 显示,随机置换特征 v_7, v_8, v_9, v_{10} 后,样本 OOB 误差率 e 变化很小,但 Prox 矩阵发生了较明显的变化. 结果表明:OOB 准确率的变化,不能够完全如实反映分类器泛化性能的变化情况,尤其是对存在大量冗余的、特征之间有很强的相关性的数据集,某些特征的随机置换并不会影响 OOB 准确率. 而 Prox 矩阵对特征的改变具有较强的敏感性.

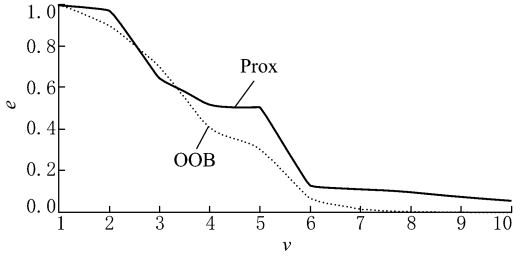

图 2　Prox 矩阵对特征敏感性分析

c. 基于 Prox 矩阵差异性的特征选择算法.

根据上述分析,Prox矩阵对RF算法的鲁棒性和对特征变化的敏感性表明,可以用其作为度量指标进行特征选择.类似RF变量重要性分析方法,采用特征随机置换的方法,提出如下的特征选择算法.

设有 l 个特征的训练样本集为 $Z=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,生成的RF为 $f=\{h_1,h_2,\cdots,h_m\}$,同时得到一个 $n\times n$ 维的Prox矩阵 $P=\{p_{ij};\ i,j=1,2,\cdots,n\}$.

a. 计算训练集中类内和类间样本的相似性比率:$C=P_s/P_d$,其中:$P_s=\sum_{i,j=1}^{l}p_{ij}(y_i=y_j)$;$P_d=\sum_{i,j=1}^{l}p_{ij}(y_i\neq y_j)$.

b. 随机改变训练集中某个变量 q 的值,即加入干扰噪声,得到新的训练集 Z^q,将 Z^q 带入已生成的RF中,得到一个新的 $n\times n$ 维的Prox矩阵 P^q.

c. 利用各个 P^q 计算加入噪声后,类内和类间样本的相似性比率 $C_i(i=1,2,\cdots,l)$,相似性比率的变化差异为 $E^q=C-C_i(v=1,2,\cdots,l)$;则 E^q 代表变量 q 对于样本相似性的影响程度.

d. 计算 E^q 的方差 $S^2=[1/(l-1)]\sum_{i=1}^{l}(E_i^q-E^q)^2$,据此估计变量 q 的重要性为 E^q/S.

2 试验与分析

2.1 实验设计

为了客观评价基于OOB准确率的特征选择算法及Prox-FS算法的性能,本文采用KNN作为基准分类器,根据其平均准确率进行评价.KNN对于距离度量比较敏感,选择恰当的特征就能取得较好的效果,且直观意义比较明显[10].两种特征选择方法的实验环境均采用R Language V2.9.0的Random Forest Package.当构建森林时,选择决策树的个数为 $m_{try}=2\ 000$,叶节点上随机分裂属性个数为 $m_{try}=[\sqrt{l}]$.对某一特征随机置换10次,取Prox矩阵的平均值作为扰动后的Prox矩阵.近邻数 k 取{1,3,5,7,9},进行0重交叉实验,并重复100次,记录每组参数(S,k)下基于欧式距离的KNN分类器的分类准确率的均值.

2.2 数据集描述

本文选取2个被广泛采用的人造数据集monk1,monk3以及其他6个UCI标准数据库的数据集来说明Prox-FS特征选择算法的性能.数据集描述见表1.

表1 数据集描述

数据集	样本数	特征数	类别
monk1	432	6	2
monk3	432	6	2
dermatology	366	34	6
parkinsons	195	22	2
ionosphere	351	34	2
SAheart	462	9	2
wine	178	13	3
zoo	101	16	7

monk1数据集中6个特征向量V与类属性的关系是:

$$\text{label}(x)=\begin{cases}1 & (v_1=v_2;v_5=1);\\ 2 & (\text{其他}).\end{cases}$$

monk3数据集中6个特征向量V与类属性的关系是:

$$\text{label}(x)=\begin{cases}1 & (v_5=3,v_4=1\ \text{或}\\ & \ v_5\neq 4,v_2\neq 3);\\ 2 & (\text{其他}).\end{cases}$$

2.3 实验结果及分析

a. monk数据集实验.由monk数据集的特点可知,monk1中与类标识相关的特征是 v_1,v_2 和 v_5.monk3中与类标识相关的特征是 v_2,v_4 和 v_5.表2列出了Prox-FS方法对monk数据集所有特征向量的重要程度降序的排序.

表2 monk数据集特征向量排序

数据集	特征					
monk1	v_5	v_1	v_2	v_3	v_6	v_4
monk3	v_2	v_5	v_4	v_1	v_6	v_3

由表2可以看出,Prox-RF算法得到的特征排序与monk数据集自身的重要特征相一致,这说明Prox-FS算法可以有效的对特征重要性进行排序.

b. UCI标准数据集实验.图3~8给出了在6个数据集上使用Prox-FS与RF变量重要性的平均预测准确率,随特征子集大小改变(按重要性递减排序)的变化曲线,其中:横轴 v 表示不同的特征子集;纵轴为平均准确率 r.

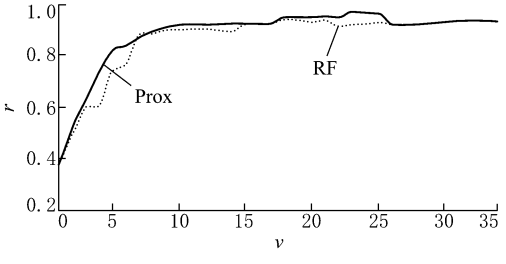

图3 两种算法在dermatology数据集上的比较

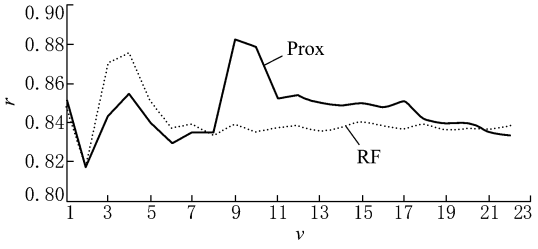

图 4　两种算法在 parkinsons 数据集上的比较

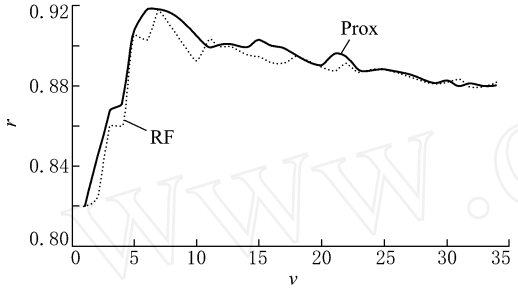

图 5　两种算法在 ionosphere 数据集上的比较

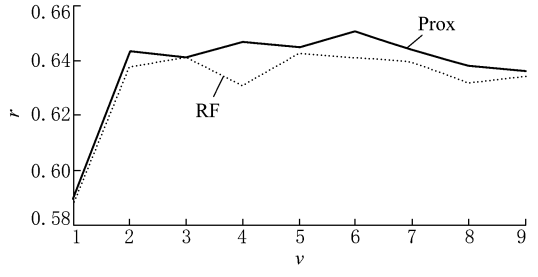

图 6　两种算法在 SAheart 数据集上的比较

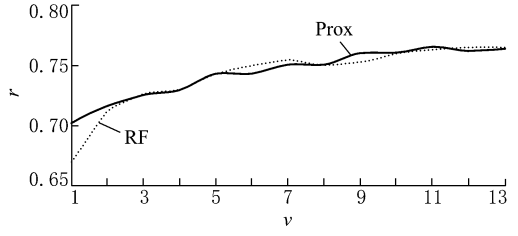

图 7　两种算法在 wine 数据集上的比较

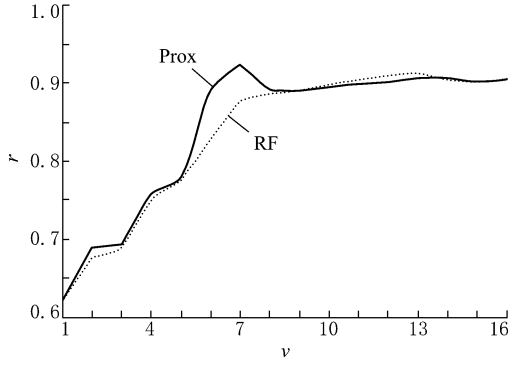

图 8　两种算法在 zoo 数据集上的比较

由图 3～8 可以看出,6 个数据集中,除了在 ionosphere 和 wine 数据集上两种方法基本持平外,在其余 4 个数据集上,本文所提 Prox-FS 特征选择算法在整体上优于基于 OOB 准确率的特征选择算法,且具有比 OOB 准确率更灵敏的特征重要性分析性能. 如何进一步提高算法的效率,及对 Proximity 矩阵进行更深入的分析,使其适用于"高维小样本"等数据集是今后继续研究的一个内容.

参　考　文　献

[1] Blum A L, Langley P. Selection of relevant features and examples in machine learning[J]. Artificial Intelligence, 1997, 97(1-2): 245-271.

[2] 陈　彬,洪家荣,王亚东. 最优特征子集选择问题[J]. 计算机学报, 1997, 20(2): 133-138.

[3] Langley P. Selection of relevant features in machine learning[C] //Proceedings of the AAAI Fall Symposium on Relevance. New Orleans: AAAI Press 1994, 1-5.

[4] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines [J]. Machine Learning, 2002, 46(1): 389-422.

[5] Xing E P, Feature selection in microarray analysis, in a practical approach to microarray data analysis[M]. Dordrecht: Kluwer Academic Publishers, 2002.

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

[7] Breiman L. Bagging Predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[8] Breiman L, Friedman J H, Olshen R A, et al. Classification and regression trees[M]. Cole: Wadsworth & Brooks, 1984.

[9] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.

[10] Hart P. The condensed nearest neighbor rule[J]. IEEE Transactions on Information Theory, 1968, 14(3): 515-516.