文章编号:1001－9383(2019)03－0037－05

随机森林算法研究综述

吕红燕，冯 倩

(河北经贸大学信息技术学院,河北 石家庄 050061)

摘 要:随机森林算法是一种基于决策树的集成学习算法,具有很高的预测准确率,对异常值和噪声具有很好的容忍度,而且不容易出现过拟合,在医学等领域具有广泛的应用。首先介绍了随机森林算法的原理和性质,然后综述了近几年来随机森林算法的改进研究及应用领域,最后对随机森林算法研究做出了总结。

关键词:随机森林;集成学习;机器学习;决策树

中图分类号:TP311 文献标识码:A

DOI:10.16191/j.cnki.hbkx.2019.03.005

A review of random forests algorithm

LV Hong-yan, FENG Qian

(*College of Information Technology, Hebei University of Economics and Business, Shijiazhuang Hebei* 050061, *China*)

Abstract: Random forest algorithm is an integrated learning algorithm based on decision tree, which has high prediction accuracy, good tolerance to outliers and noise, and is not easy to overfit, and has a wide range of applications in medicine and other fields. This paper first introduces the principle and properties of random forest algorithm, then summarizes the improvement of random forest algorithm and its application fields in recent years, and finally summarizes the research of random forest algorithm.

Keywords: Random forest; Integrated learning; Machinelearning; The decision tree

引言

机器学习算法中的有监督学习无非就是解决分类问题和回归问题,其中解决分类问题的算法有很多,例如朴素贝叶斯算法(Naive Bayesian)、支持向量机算法(Support Vector Machine)、决策树算法(Decision Tree)等。显然这些都是单个的分类器,单个的分类器很容易出现过拟合的问题,而且在对其进行性能提升时会出现瓶颈,因此集成学习算法运应而生。集成学习算法中集成方法主要有两种:Boosting 和 Bagging(boostrap aggregating)。其中 Boosting 算法有很多,最具代表性的应当是 AdaBoost 算法;Bagging 是通过结合几个模型降低泛

收稿日期:2019－05－21

作者简介:吕红燕(1994－),女,河北沧州人,硕士研究生,研究方向:行为分析和知识建模.

化误差的技术，随机森林(Random Forests)是 Bagging 集成方法中最具有代表性的算法，该算法是 2001 年由 Leo Breiman 将 Bagging 集成学习理论与随机子空间方法相结合，提出的一种机器学习算法。本文将以随机森林算法的研究作为重点进行综述。

1 随机森林算法简介

1.1 决策树

决策树[1]是一类常见的机器学习算法，是基于树结构来进行决策的一种算法。决策树算法有 ID3、C4.5、CRAT、SLIQ 等[2-3]，其中 ID3 是由 Quinlan 提出的，C4.5 和 CART 都是从 ID3 决策树算法中衍生而来的，SLIQ 算法则是在 C4.5 决策树分类算法的实现方法上进行了改进，C4.5 是按照深度优先策略构造树的，而 SLIQ 是按照广度优先策略来构造树结构的。一棵决策树的生成过程主要有 3 个部分，即特征选择、决策树生成和剪枝。其中最关键的问题是特征选择，不同的分裂标准对决策树的泛化误差有很大的影响。ID3 决策树算法是根据信息论的信息增益来进行评估和特征选择的，C4.5 决策树算法是用信息增益率来选择特征的，CART 决策树算法采用的是 Gini 指数来进行选择的。

决策树算法适用于离散型数据，能够提取出列数据中蕴含的规则，不需要先验知识，比神经网络等方法更容易解释，在解决分类问题时，决策树算法具有计算复杂度不高，便于使用且高效的优点。但是决策树算法在处理缺失数据时很困难，此外，可能会对样本空间过度分割导致过拟合的问题。通过剪枝的方法避免决策树的过拟合问题又会提高算法的复杂性。所以决策树算法的性能提升有一定的局限性。

1.2 Bagging 集成学习方法

集成学习的本质是通过训练若干个弱学习器，经过一定的结合策略最终形成一个强学习器，集成学习方法的原理如图 1 所示。集成学习方法主要分为并行和串行两种方法。Bagging[4]是并行式集成学习方法的典型代表，直接基于自助采样法。对训练集合采用有放回的随机抽样，所以每轮的分训练集由训练集中 N 个样本构成，某个训练样本在一轮训练集中可以出现多次或者根本就不出现。将随机抽取的子集放到算法中训练，计算预测函数，T 轮循环后得到一个结果集。最终的预测函数对分类问题采用投票方式，对回归问题采用简单平均方法判别。

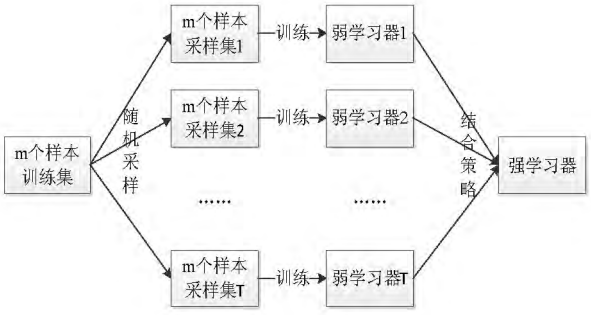

图 1 集成学习方法的原理图

1.3　随机森林算法

随机森林在以决策树为基学习器构建 Bagging 集成的基础上,进一步在决策树的训练过程中引入随机属性的选择。随机森林算法简单、易于实现、计算开销小,在很多现实任务中展现出强大的性能。FernandezDelgado 等人[5]通过大量实验在 121 个 UCI 数据集上比较了 179 种分类算法的分类性能,实验结果表明,随机森林算法的分类性能是最优秀的。

随机森林分类是由很多决策树分类模型组成的组合分类模型,每个决策树分类模型都有一票投票权来选择最优的分类结果。随机森林分类的基本思想:首先,利用 bootstrap 抽样从原始训练集抽取 k 个样本,每个样本的样本容量都与原始训练集一样;然后,对 k 个样本分别建立 k 个决策树模型,得到 k 种分类结果;最后,根据 k 种分类结果对每个记录进行投票表决决定其最终分类。其示意图如图 2 所示。

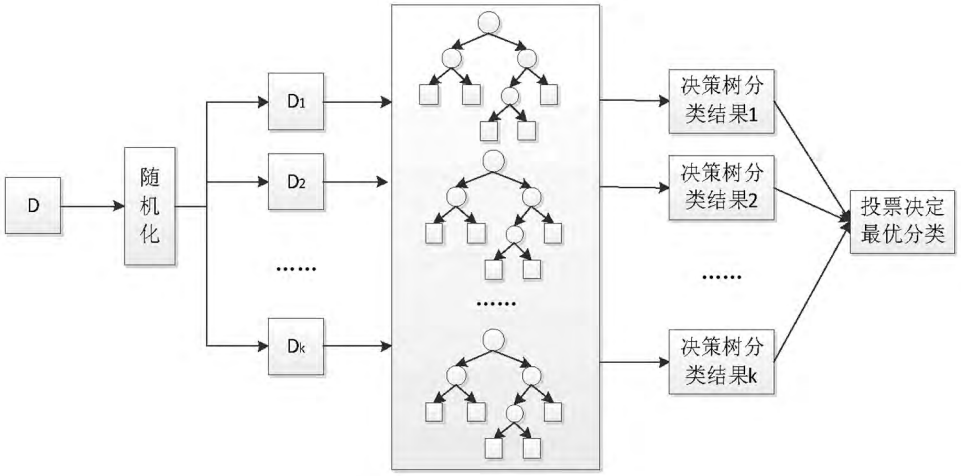

图 2　随机森林分类示意图

2　随机森林算法改进研究综述

随机森林算法虽然在分类精度、泛化误差等性能方面比决策树算法有较大的提升,也得到了广泛的应用,但是随着研究的深入,随机森林算法自身存在的问题也慢慢暴露出来,主要表现在三个方面:第一,不能很好地处理非平衡数据;第二,对连续性变量的处理还需要进行离散化;第三,分类精度还需要进一步提高。

2.1　处理非平衡数据上的改进

随机森林算法不能很好地处理非平衡数据的原因主要是因为随进森林构建时的训练集是随机选取的,加剧了数据集的不平衡性。黄衍等人[6]使用 20 个 UCI 数据集比较了随机森林算法和支持向量机算法在不平衡分类中的分类效果,实验表明,随机森林在不平衡数据上的分类性能明显不如支持向量机。

不少学者在这方面做了相关研究,吴琼[7]等人在使用随机森林算法对非平衡数据进行分类时先使用 NCL(Neighborhood Cleaning Rule)技术对数据进行了处理,再应用随机森林算法,提高了分类的准确率。王雪[8]在处理高维不平衡数据时采用了欠抽样和过抽样处理使原始数据集达到平衡状态。钟龙申[9]提出一种 KSMOTE 算法,提高了随机森林算法的分类性

能。徐少成[10]在数据方面提出了一种改善数据平衡问题的优化的 SMOTE 算法——E-SMOTE 算法，平衡后的数据有效缓解了数据的不平衡性对模型的影响。马海荣、程新文[11]在解决传统随机森林模型随机抽取样本时训练样本集中包含不同类别样本数不平衡问题时，采用的方法是随机抽取等量的少数类与多数类样本构建训练样本集进行随机森林建模，然后根据投票熵与基于样本特征参数的广义欧几里得距离，逐步添加具有最大投票熵的样本到训练样本集，有效解决了数据不平衡的问题。赵锦阳、卢会国等人[12]针对随机森林算法在不平衡数据集上表现的分类性能差的问题，提出了一种新的过采样方法：SCSMOTE 算法，并且在 UCI 数据集上对比实验表明该算法有效提高了随机森林在非平衡数据集上的分类性能。总体看来，在解决随机森林算法在非平衡数据上分类性能差的问题上，学者们采取的方法基本是对数据集进行预处理，使得数据集处于一个相对平衡的状态。

2.2 处理连续性变量的改进

传统的随机森林中，如果存在连续性变量的做法是将其分成不同的区间，即离散化。但是这种方法会使得算法在分析计算节点分裂标准是花费大量的时间，会影响算法的执行速度。目前连续变量离散化的方法有很多，最主要的方法是基于统计学思想的 CHI2 相关算法。曹正凤[13]在使用随机森林处理连续性变量时借鉴和改进了 CHI2 算法，提出了 COR_CHI2 算法，提高了随机森林算法的执行效率。

2.3 在提高分类精度上的改进

任何一个分类算法的分类精度越高越好，也就是准确度(accuracy of measurement)。所以，分类精度的提高是分类算法优化研究中永恒的主题。随机森林分类算法的分类精度虽然相对较高，但是在不同数据集上的分类精度还是存在一些问题的。许多学者对随机森林算法进行了广泛的研究，并且取得了显著的成果。对随机森林算法的改进主要在以下几个方面：①对数据集进行预处理工作；②改进生成的决策树算法；③对生成的决策树进行筛选；④改进投票方式。前面提到的随机森林对非平衡数据和连续性数据的处理就是对数据集进行了预处理工作。曹正凤等人[14]将 C4.5 决策树算法和 CART 决策树算法混合为一个算法，使用混合后的算法生成随机森林算法，提高了随机森林的精确度。王日升等人[15]将随机森林模型中的决策树按 AUC(area under curve)值进行降序排序，选取 AUC 值高的决策树，然后计算这些决策树的相似度，并生成相似度矩阵，聚类之后选取每类中 AUC 值最高的组成随机森林模型。实现表明，改进后的随机森林算法在分类精度上有所提高。Paul 等人[16]提出了一种改进的随机森林分类器，以最小树数进行分类，根据特征的重要性限制随机森林中决策树的数量。在不同数据集上实验表明萍爵分类误差有显著降低。王诚等人[17]提出了一种基于决策树聚类的改进随机森林算法，取出分类精度低和相似性高的决策树，通过实验表明该算法在集成准确率和分类效率上高于传统的随机森林算法。

3 随机森林算法的应用

由于随机森林算法的综合性能较好，所以在很多领域都有广泛的应用。许允之[18]把随机森林算法应用到环境保护中，用其预测徐州雾霾情况，最后分析和阐述了徐州对雾霾的治理措施。Li 等人[19]把随机森林算法应用到了推荐系统中，提出了一种基于改进随机森林算法的多维上下文感知推荐方法，实验结果表明该方法可以降低平均绝对误差和均方根误差。Jonny

Evans 等人[20]利用基于上下文的随机森林算法预测道路交通状况。de Santana Felipe Bachion 等人[21]利用随机森林算法和红外光谱进行食品掺假检测。

综上所述,随机森林算法是一种分类精度较高,效率也较高的算法,而且其理论和方法的研究都已经比较成熟,在很多领域都有所应用并且效果较好。

参考文献:

[1] 唐华松,姚耀文. 数据挖掘中决策树算法的探讨[J]. 计算机应用研究,2001(08):18－19＋22.

[2] Quinlan J R. Induction of decision trees[J]. Machine Learning,1986,1(1):81－106.

[3] Quinlan J R. C4. 5:programs for machine learning[J]. 1992.

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

[5] Fernandez-Delgado M,Cernadas E,Barro S,et al. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems[J]. Journal of Machine Learning Research,2014,15:3133－3181.

[6] 黄衍,查伟雄. 随机森林与支持向量机分类性能比较[J]. 软件,2012,33(06):107－110.

[7] 吴琼,李运田,郑献卫. 面向非平衡训练集分类的随机森林算法优化[J]. 工业控制计算机,2013,26(7):89－90.

[8] 王雪. 面向高维不平衡数据的随机森林算法及其并行化研究[D]. 辽宁大学,2016.

[9] 钟龙申. 随机森林算法处理不平衡数据的改进及其并行化[D]. 广东工业大学,2016.

[10] 徐少成. 基于随机森林的高维不平衡数据分类方法研究[D]. 太原理工大学,2018.

[11] 马海荣,程新文. 一种处理非平衡数据集的优化随机森林分类方法[J]. 微电子学与计算机,2018,35(11):28－32.

[12] 赵锦阳,卢会国,蒋娟萍,袁培培,柳学丽. 一种非平衡数据分类的过采样随机森林算法[J]. 计算机应用与软件,2019,36(04):255－261＋316.

[13] 曹正凤. 随机森林算法优化研究[D]. 首都经济贸易大学,2014.

[14] 曹正凤,谢邦昌,纪宏. 一种随机森林的混合算法[J]. 统计与决策,2014(04):7－9.

[15] 王日升,谢红薇,安建成. 基于分类精度和相关性的随机森林算法改进[J]. 科学技术与工程,2017,17(20):67－72.

[16] Paul A,Mukherjee D P,Das P,et al. Improved Random Forest for Classification[J]. IEEE Transactions on Image Processing,2018:1－1.

[17] 王诚,王凯. 一种基于聚类约简决策树的改进随机森林算法[J/OL]. 南京邮电大学学报(自然科学版),2019(03):91－97[2019－06－30]. https://doi. org/10. 14132/j. cnki. 1673－5439. 2019. 03. 013.

[18] 许允之. 基于随机森林算法的徐州雾霾回归预测模型[A].《环境工程》编委会、工业建筑杂志社有限公司.《环境工程》2019 年全国学术年会论文集[C].《环境工程》编委会、工业建筑杂志社有限公司:《环境工程》编辑部,2019:6.

[19] Li X,Wang Z,Wang L,et al. A Multi-Dimensional Context-Aware Recommendation Approach Based on Improved Random Forest Algorithm[J]. IEEE Access,2018,6:1－1.

[20] Jonny Evans,Ben Waterson,Andrew Hamilton. Forecasting road traffic conditions using a context-based random forest algorithm[J]. Transportation Planning and Technology,2019,42(6).

[21] de Santana Felipe Bachion,Borges Neto Waldomiro,Poppi Ronei J. Random forest as one-class classifier and infrared spectroscopy for food adulteration detection[J]. Food chemistry,2019,293.