机器学习随机森林算法的应用现状

文/杭琦　杨敬辉

摘要

随机森林（RF）是机器学习算法中的一种组合分类器，也是集成学习的代表性算法之一。它通过bagging算法集成多个决策树并以投票的形式输出结果，在学术界和工业界均取得了很好的评价。本文将具体介绍随机森林算法的构建过程，总结随机森林算法在性能改进、性能指标方面的研究，对目前随机森林已经有的理论和应用研究做一个系统的总结和整理，以利于后续的算法优化研究。

【关键词】机器学习　集成学习　随机森林

机器学习算法主要解决的是分类和聚类的问题。分类问题是根据用户的分类数据得到预测的分类结果。根据分类器的个数，分类器又分为单分类器和多分类器。例如决策树、贝叶斯都是传统单分类算法。这些传统的机器学习算法在一定程度上都促进了分类学习的发展，但由于单分类器有其自身的限制，容易产生过拟合等现象。故学者们提出集成多个分类器形成组合分类器，把一个学习问题分解到各个子学习器内，让其一起学习。多分类器的分类思想起源于集成学习，Boosting 和 Bagging 是最早将集成学习思想应用到机器学习分类算法里中两种算法。随着集成学习的发展，Tin Kam Ho 在 1995 年提出了随机决策森林的思想，1998 年，他又提出了新的随机子空间的集成方法，Breiman 根据随机子空间的思想在 2001 提出了随机森林算法，从理论和实践两方面做了系统的阐述，自此随机森林算法成为机器学习领域中的一个具有代表性的集成学习的方法。

本篇文章第一节针对随机森林算法构建过程进行简单介绍；第二节介绍随机森林在性能改进方面的研究；第三节针对随机森林的性能指标进行研究总结；最后总结全文。

1 随机森林算法的构建过程

随机森林算法是一种集成分类模型，它的构建过程主要由三个方面构成，训练集的生成、决策树的构建和算法的产生。要构建随机森林首先要生成一个规模大小为 N 的随机森林，就需要有 N 颗树，因此需要 N 组训练集。故首先我们需要从原始数据中通过抽样产生训练集。通过 Bagging 算法从原始数据集中抽取 N 个样本。每个样本都会生产一个决策树，且生成的决策树不需要做剪枝处理，从而建立起 N 棵决策树形成森林。随机森林生成过程中涉及到如下三个评估过程：

（1）指定 m 值，由于在每棵决策树分裂的过程中，不是样本中全部 K 个特征属性都参与分裂，而是从中随机抽取 m 个变量，同时分裂过程中特征属性的选择需满足节点不纯度最小原则。

（2）应用 Bagging 随机取样法在原数据集中有放回地随机抽取 k 个样本集，组成 k 棵决策树；

（3）根据 k 个决策树组成的随机森林对待分类样本进行分类或预测，分类的结果由单颗决策树的分类结果投票决定。

从随机森林三个评估过程中可以看出。随机森林的构建过程中掺入了随机性，从而降低了随机森林过拟合现象的产生。

2 随机森林算法优化方法研究

基于集成学习的随机森林算法从根源上改善了决策树容易过拟合的特性。但是该算法在算法处理不同类型数据集特别是不平衡数据集和算法分类精度的方面，还需要一定程度的改进。因此国内外的学者专家们就随机森林算法的优化方面提出了很多的改进的方法，细分起来，它们可以分成以下三个主要的方面。

2.1 结合数据预处理对随机森林算法进行优化

不平衡数据集的分类问题是当前机器学习领域的一大挑战。故针对随机森林处理不平衡数据集上的分类问题上，学者专家们将数据预处理融入到随机森林算法优化的研究中来，通过数据预处理，随机森林的性能得到了一定的提升。

文献 [4] 提出代价敏感随机森林算法，在随机森林算法中引入代价敏感学习，让分类器更偏向少数类，使得总的误分类代价最小化。文献 [5] 首次提出对原始数据进行 NCL（Neighborhood Cleaning Rule）处理，并将处理过的数据结合随机森林算法进行分类，实验表明经过 NCL 技术改进后的随机森林算法拥有更好的分类精度。文献 [6] 提出了分层抽样的随机森林算法，并且在节点分裂处采用支持向量机算法作为算法的基分类器，结果表明经过改进的随机森林算法在非平衡数据的处理效果比传统的随机森林、过采样支持向量机、欠采样支持向量机的都要好。

2.2 针对随机森林算法构建过程的优化

针对算法自身构建过程的改进主要表现在降低泛化误差，减少每颗决策树之间的相关性。

由于传统随机森林算法中各个决策树的之间的权重相同，故修改决策树之间权重的思想被广泛的用于随机森林的改进。Li，Wang[7] 等人根据袋外数据误分率进行权重设置，用正确的分类与随机森林分类器的结果进行比较，统计随机森林分类器分类错误的数目。雍凯 [8] 利用卡方检验进行特征的相关性评估，依据评估的结果进行随机特征选择，该方法可以很好的降低随机森林泛化误差的上界，进而提高整体的分类精度。孙丽丽 [9] 等人根据由聚类数据构建的多棵决策树构成的随机森林来进行分类器的加权集成，通过加权集成可以很好的降低数据集的复杂性，提高整体的分类效率和分类准确度。

2.3 引入新算法进行随机森林的优化

Breiman 根据随机子空间的思想在 2001 提出了随机森林算法。从本质上讲，该算法是 Bagging 方法和 Random Subspace 方法的组合。近几年来对于随机森林的改进方法的研究大部

● 本项目受上海第二工业大学机械工程学科建设资助（XXKZD1603）。

表 1：混淆矩阵

ConfusionMatrix	Classifiedpositive	Classifiednegative
Positive	TP	FN
Negative	FP	TN

分在组合算法上，通过将优秀的算法融入到随机森林算法中，从而提升分类精度。

旋转森林算法 (Rotation Forests)[10] 中引入了主成分分析算法进行特征向量的变换，通过把原始数据集上的原始向量通过坐标变换旋转到主成分所在的方向，再进行随机森林的构建。霍夫森林算法（Hough forests）[11] 将霍夫变换引入到随机森林的投票过程中，对随机森林的投票机制进行了优化。马景义 [12] 等我国学者们将 Adaboost 算法与随机森林算法进行组合，提出了一种改进的随机森林算法算法—拟自适应分类随机森林算法，此算法不用区分数据集，通过发挥两种算法各自的优势，得到了较好的分类效果。

从上文的三种优化方法来看，对于随机森林算法分类性能的提升，第一种改进方法主要侧重于对于不平衡数据的优化研究上；第二种改进方法主要集中于各种组合算法的研究上，这些组合算法一般都被用于某个特定的问题上；而第三个种优化方式主要集中在算法本身的改进上，在权重的优化方面改进较多，这类算法具有一定的通用性，可以在不同的领域中使用。

3 随机森林算法的性能指标研究

随机森林分类性能受外部因素和内部因素的共同影响，从内部因素来看，一般从每棵决策树的最大树深度、决策树的分类强度和决策树之间的相关性来考虑。从外部因素看，主要来自原始数据本身的分布情况，包括正负样本的分类，样本的规模等情况。评价随机森林性能的指标一般有两种 : 分类效果指标和泛化误差。

3.1 分类效果指标

随机森林算法的应用场景最多的还是出现在预测和分类模型中，表 1 中的混淆矩阵是二分类中经常用到的评估分类效果的指标。

其中 TP 指被模型预测为正的正样本数量; TN 指的是被模型预测为负的负样本数量；FP 指被模型预测为正的负样本数 ;FN 指的是被模型预测出来为负的正样本数。

精确率（Precision），表示预测为正例的样本中，真正为正例的比例计算公式为：

$$\text{Precesion}=\frac{TP}{TP+FP}$$

分类准确率 (Accuracy)，表示分类模型总体的分类精度，计算公式为：

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

3.2 泛化误差

泛化能力（generalization ability）是指机器学习算法对新鲜样本的适应能力。即对于具有相同分布规律训练集以外的数据，该模型也能做出正确的判断。在很多工业生产的应用场景中，我们通常用泛化误差（Generalization Error）来评估机器学习算法的泛化能力，如果泛化误差越大，那么该模型学习性能越差，反之则性能越好。泛化误差从理论上来说可以通过公式直接计算出来的。但是从实际应用来看我们无法获得准确的样本分布情况和样本的期望输出。目前用来估计分类器的泛化误差的方法主要有两种，一种是分析模型 (Analytical Model) 还有一种是交叉验证 (Cross-Validation, CV)。分析模型对于简单的线性分析是比较有效的，但由于其难以对随机森林的有效参数做出合理估计，所以对于非线性等复杂的问题难以突出其优点。交叉验证是通过把训练数据分成了训练集和测试集，用训练集来训练算法，再用测试集来验证算法，从而通过验证集来估计泛化误差。而 OOB 估计是随机森林算法的一种比较好的估计泛化误差的方法。在构建决策树时需要对训练集进行随机且有放回地抽取，故对于随机森林模型中的初始训练集来说总会有一些原始数据没有参加模型的训练，而这些没有参加模型训练的样本就是OBB数据。Breiman 已经在论文中用实例表明 OOB 估计与同样误差大小的测试集有着相同的分类精度，OBB 估计可以作为随机森林泛化误差的一个无偏估计。

4 结语

在如今，机器学习算法是被很多学者专家们追捧的学习方法，随机森林也是在其中孕育而生的。随机森林算法是集成学习中具有代表性的一个算法，它简单高效、应用广泛，在金融学、医学、生物学等众多应用领域均取得了很好的成绩。故本文对随机森林构建过程做了研究，还通过其性能改进和性能指标两方面进行了总结。但作为学术界和工业界均应用很广的一个算法，我们还需要考虑在数据量日益增大的复杂分类任务中，在如何有效提升模型复杂度，如何处理非平衡数据、连续性数据以及如何提升算法的分类精度这些问题上都值得我们进行深入的探讨。

参考文献

[1]T.K.Ho.RandomDecision Forest[J].In Proceedings of the3rd International Conference on Document Analysis and Recognition. Montreal, Canada, 1995,8:278-282.

[2]T. K. Ho, he Random Subspace Method

>> 下转 127 页

新媒体时代计算机图形图像处理技术在传媒中的应用

文/方芳

摘要

在科学技术不断进步的发展背景下，数字图像处理技术迎来了空前发展，与此同时也提高了图像技术在传媒行业中的价值。随着新媒体时代的到来，数字图像处理技术的应用涉及到了各行各业，对图像传媒行业的影响也是不容忽视。数字图像处理技术在很大程度上推动了图像传媒行业的可持续发展。数字图像处理技术主要包括数字图像编码压缩技术、分割技术、增强复原技术，本文通过对数字图像处理关键技术的研究，旨在让从事图像处理的同行更深入的了解此项技术，为以后的研究提供理论基础。

【关键词】新媒体时代 计算机图像 处理技术 应用

在传统媒体中，图文结合一直都是报刊的特色。只有文字的报刊，可读性并不高，当一份报刊图文并茂的时候，才会受阅读者的喜爱。图片已不再是单单在版面上起点缀、装饰作用，许多图片题材重大、现场感强、形象直观而在新闻报道中唱主角、挑大梁。

1 图形图形处理技术概念

图像处理（image processing），是指用计算机对图像进行分析，以达到所需结果的技术，又称影像处理。图像处理一般指数字图像处理，数字图像是指用工业相机、摄像机、扫描仪等设备经过拍摄得到的一个大的二维数组，该数组的元素称为像素，其值称为灰度值。图像处理技术的一般包括图像压缩，增强和复原，匹配、描述和识别 3 个部分。 常见的系统有康耐视系统、图智能系统等，目前是正在逐渐兴起的技术。21 世纪是充满信息的时代，图像作为人类感知世界的视觉基础，是人类获取信息、表达信息和传递信息的重要手段。数字图像处理，即用计算机对图像进行处理，其发展历史并不长。

2 图形图像处理技术在传媒中的应用

随着新媒体时代的到来，图片的价值作用越来越凸显。我们利用计算机技术，将更多的文字描述，制作成一张张鲜活的图片，不仅让读者的视觉得到满足，也更加有效率的传播了信息，图片给人留下的印象会比文字更加深刻，这也是为什么图片处理技术在新时代传媒中的作用价值会越来越高的原因。

3 数字图像处理常用技术

数字图像处理是采用计算机技术将图像进行压缩、编码、增强等数字化处理，从而将低画质的图像转变为高质量的图像的过程。下面介绍几种数字图像处理常用方法。

3.1 数字图像编码压缩技术

数字图像编码压缩技术在图像网络传输方面应用最多，这样可以方便图像的传输，减少图像占用空间，节省资源。压缩需要考虑图像的失真程度，根据压缩的可逆性，可以将压缩分为可逆压缩和不可逆压缩。编码是压缩的常用方式。常见的图像压缩编码格式有 BMP、Jpeg、Gif、TGA、FPX、PNG、EPS 等。下面对各种编码格式下的压缩效果进行比较，方法将同一张大小为 1Mb 的 TGA 图片进行压

<< 上接 126 页

for Constructing Decision Forests. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998,20(8):832-844.

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

[4]houQ, Zhou H, Li T. Cost-sensitive feature selection using random forest: Selecting low-cost subsetsof informativefeatures[J]. Knowledge-Based Systems, 2015:S0950705115004372.

[5] 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 201213,26(07):89-90

[6]Wu Q, Ye Y, Zhang H, et al. ForesTexter: An efficient random forest algorithm for imbalanced text categorization[J]. Knowledge-Based Systems, 2014, 67(3):105-116.

[7]Li H B, Wang W, Ding H W, et al. Trees Weighting Random Forest Method for Classifying High-Dimensional Noisy Data[C]. IEEE International Conference on E-business Engineering. IEEE, 2011.

[8] 雍凯．随机森林的特征选择和模型优化算法研究 [D]．哈尔滨工业大学，2008.

[9] 孙丽丽．基于属性组合的随机森林 [D]．河北大学，2011.

[10]Rodriguez J J, Kuncheva L I, Alonso C J. Rotation Forest: A New Classifier Ensemble Method[J]. IEEE Trans Pattern Anal Mach Intell, 2006, 28(10):1619-1630.

[11]LempitskyV. Hough Forests for Object Detection, Tracking, and Action Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2011,33(11):2188-202.

[12] 马景义，吴喜之，谢邦昌．拟自适应分类随机森林算法 [J]．数理统计与管理，2010,29(05):805-811.

[13]Cortes C. Prediction of Generalization Ability in Learning Machines[J]. 1995.

[14] 刘凯．随机森林自适应特征选择和参数优化算法研究 [D]．长春工业大学，2018.

作者简介

杨敬辉，女，博士学历。教授。研究方向为数据分析，智能制造。

杭琦，硕士研究生。研究方向为环境工程。

作者单位

上海第二工业大学　上海市　201209