Computer Science and Application 计算机科学与应用, 2020, 10(9), 1541-1548
Published Online September 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.109162

Hans汉斯

用于无人机巡线的图像分类模型选择算法研究

金明磊[1]，李　明[1]，赵文璇[2]

[1]天津航天中为数据系统科技有限公司，天津
[2]国家海洋技术中心，天津
Email: 15900364481@126.com

收稿日期：2020年8月15日；录用日期：2020年8月26日；发布日期：2020年9月2日

摘　要

随机森林算法作为经典的分类算法，应用广泛，分类的准确度高。但在分类的过程之中，各个决策树的分类性能和两两决策树之间的差异性是影响最终分类效果的两个重要因素，当部分决策树有相似的错误分类情况，在最终利用决策树的结果进行投票时，将降低模型最终的分类效果。针对该问题，本文将误差矩阵引入分类树的相似性度量当中。该方法考虑了不同类别的树的数量、分类正确错误的情况，以便选出相似度弱的决策树，然后，剔除分类能力差的决策树，最终选择出分类能力强的分类器集合。实验结果显示，本文提出的方法在3类数据集中，平均分类正确率高于原算法，且稳定性更高。

关键词

集成分类器，随机森林，误差矩阵

Research on Algorithm of Image Classification Model Selection for UAV Patrol

Minglei Jin[1], Ming Li[1], Wenxuan Zhao[2]

[1]Tianjin Zhongwei Aerospace Data System Technology Co. Ltd., Tianjin
[2]National Ocean Technology Center, Tianjin
Email: 15900364481@126.com

Received: Aug. 15th, 2020; accepted: Aug. 26th, 2020; published: Sep. 2nd, 2020

Abstract

As a classic classification algorithm, random forest algorithm is widely used and has high classification accuracy. However, in the process of classification, the classification performance of each decision tree and the difference between two decision trees are two important factors that affect

文章引用: 金明磊, 李明, 赵文璇. 用于无人机巡线的图像分类模型选择算法研究[J]. 计算机科学与应用, 2020, 10(9): 1541-1548. DOI: 10.12677/csa.2020.109162

the final classification effect. When some decision trees have similar misclassifications, and they are used in the final voting on the results of the decision tree, the final classification effect of the model will be reduced. Aiming at this problem, this paper proposes a method for measuring the similarity of decision trees based on confusion Matrix. This method takes into account the number of different categories of trees and the correct and incorrect classification, in order to select decision trees with weak similarity, and then remove the decision trees with poor classification results, and finally complete the model selection of random forest. Experimental results show that the method proposed in this paper has a higher average classification accuracy rate and higher stability in the three types of datasets.

Keywords

Integrated Classifier, Random Forest, Confusion Matrix

1. 引言

随着无人机电力线路巡检的发展，无人机拍摄了大量的图像信息，数据将增长非常快，目前已经发展到了几个 PB 的海量数据。为了便于处理数据，各种分类算法层出不穷。其中，随机森林算法是以分类集成思想为基础的回归模型。该算法被应用于气象分析[1]、医学[2]、大数据推荐[3]等。由于随机森林良好的分类能力，也被用来进行数据的处理，并应用于分布式当中[4]-[9]。该算法被进行多次改进。如通过将新的理论引入随机森林，得到算法效果的提升。文献[10] [11]将随机森林算法与 Hough Transform 相结合，应用于目标检测，效果较好。文献[12]把 survival forests 与随机森林相结合，提升了算法的性能。谢晓东[13]等利用梯度提升算法森林模型进行了改进，提高了模型的分类准确性。魏正涛[14]通过对抽样结果增加约束条件来改进重抽样方法，加强了算法的分算类能力。王诚[15]等对随机森林算法中存在的面对特征纬度高且不平衡的数据时，算法性能低下的问题提出了改进的算法。该算法先对数据集的特征按照正负类分类能力赋予不同的权值，然后删除冗余的低权值的特征值，得到性能良好的特征子集构造随机森林。文献[16]使用聚类的方法，将效果良好的分类器进行聚合；文献[17]将基分类器进行划分，选择出效果良好的分类器。通过筛选出来的决策树进行最终的投票过程中，如果各个决策树的相似性过高，决策树的种类过于单一，那么最终的分类效果会变差。同时，现有的决策树选择策略带来的计算量比较复杂，分类效果欠佳。为了解决该问题，本文将误差矩阵引入随机森林算法中，删选出种类更多的决策树，提升了随机森林算法的分类能力。

2. 集成学习

集成分类器的差异性度量

集成学习，即集成分类器，指的是通过构建若干分类器，然后用某种方法将这些分类器的分类结果结合起来进行学习，完成任务。

集成分类器有两种方式进行分类。第一种是选择不具有强依赖关系的分类器进行学习，该方法将分类器进行综合和分析较为困难。第二种是选择有依赖关系的分类器进行学习，对于大多数分类方法都倾向于该方法。对于总的分类器的错误率和单分类器错误率之间的关系[18]，如公式(1)所示。

$$E_{total} = E_{bay} + \frac{1+\rho(N-1)}{N} E_{each} \tag{1}$$

其中，E_{total} 及 E_{each} 分别代表总的错误率和单分类器的错误率，ρ 表示单分类器的错误相关性，N 是总分类器的规模，E_{bay} 是基于已知分布和 Bayes 规律的分类错误率。

对于差异性的度量有两种方法。第一种是非成对差异性度量，即直接计算集成系统的差异性值。第二种是成对差异性度量，通过计算每一对分类器的差异性值，然后用平均值衡量总的差异值。常用的方式[19] [20]有分歧度量，双误度量、评判间一致度[21]等。本文采用的就是第二种方法。

3. 随机森林算法

随机森林算法具体的步骤如下所示。

1) 从训练集中采用 bootstrap 法，即自助抽样法，有放回地抽取若干样本，作为一个训练子集。

2) 对于训练子集，从特征集中无放回地随机抽取若干特征，作为决策树的每个节点的分裂的依据。

3) 重复步骤 1)和步骤 2)，得到若干训练子集，并生产若干决策树，将决策树组合起来，形成随机森林。

4) 将测试集的样本输入随机森林中，让每个决策树对样本进行决策，得到结果后，采用投票方法对结果投票，得到样本的分类结果。

5) 重复步骤 4)，直到测试集分类完成。

4. 随机森林模型选择

4.1. 本文提出的模型选择方法

本文提出一种基于误差矩阵的随机森林改进方法，主要思想在于将误差矩阵引入决策树的相似性度量方法中，以删选出合适的分类树。随机森林算法的总流程如图 1 所示，本文重点是利用误差矩阵对决策树就进行删选。

4.1.1. 基于误差矩阵判断分类树相似性

判断两棵分类树相似度的方法通常主要有两类，第一种是根据两棵树之间的结构的差异性来判断相似性，常用的方法是利用一些算法[22]直接判断结构的差异性，通常该方法的计算量较大。第二种是通过分类树的分类结果来判断相似性。如果两棵树的分类准确率相近则认为是相似的，或者两棵树对样本集进行分类时，将其分在同一类，也认为是相似的。本文采用的是第二种方法，即将分类结果中类别之间的关系进行分析，在分类树的误差矩阵上进行相似性的判别。

误差矩阵常作为分类结果的可视化工具[16]，在监督学习中应用广泛。对于误差矩阵，每一列代表数据的预测类别，每一列的总数代表预测为该类别的样本的数目，每一行代表数据的真实归属类别。每一行的总数表示该类别的数目。$X=\{x_1,x_2,\cdots,x_N\}$ 表示 N 个样本数据，$Y=\{y_1,y_2,\cdots,y_M\}$ 表示 M 种分类的类别，通过矩阵 $\boldsymbol{CN}$ 表示 N 个样本数据在分类之后的结果，如公式(2)所示。

$$\boldsymbol{CN}=\begin{bmatrix} cn_{11} & \cdots & cn_{1m} \\ \vdots & cn_{ii} & \vdots \\ cn_{m1} & \ldots & cn_{mm} \end{bmatrix} \tag{2}$$

式中 cn_{ij} 表示样本数据 X 中真实类别为 i 的数据被分为类别 j 的数据的总数量。显然 cn_{ii} 表示的是类别为 i 的数据被正确分类的数量。

本文使用矩阵的距离测度和向量夹角作为两棵树的相似性度量。当分类树类似时，则矩阵就相接近。同理，当矩阵距离较远，则分类树的差距就较大。

差值矩阵 $\boldsymbol{DCN}^{(i,j)}$ 是两个误差矩阵 $CN^{(i)}$ 和 $\boldsymbol{CN}^{(j)}$ 之差(i, j 代表两棵决策树)。 $\boldsymbol{DCN}^{(i,j)}$ 的大小是 $M \times M$ ，如公式(3)所示。

$$\boldsymbol{DCN}^{(i,j)} = \boldsymbol{CN}^{(i)} - \boldsymbol{CN}^{(j)} = \begin{bmatrix} cn_{11}^{(i)} - cn_{11}^{(j)} & cn_{12}^{(i)} - cn_{12}^{(j)} & \cdots & cn_{1N}^{(i)} - cn_{1M}^{(j)} \\ cn_{21}^{(i)} - cn_{21}^{(j)} & cn_{22}^{(i)} - cn_{22}^{(j)} & \cdots & cn_{2N}^{(i)} - cn_{2M}^{(j)} \\ \vdots & \vdots & \ddots & \vdots \\ cn_{M1}^{(i)} - cn_{M1}^{(j)} & cn_{M2}^{(i)} - cn_{M2}^{(j)} & \cdots & cn_{MM}^{(i)} - cn_{MM}^{(j)} \end{bmatrix} \tag{3}$$

当不同类别的样本的数量有较大的差距时，数量大的类别会影响到矩阵距离的计算，最终使随机森林分类器更多地趋于多值类别。因此，考虑到这个因素，本文使对差值矩阵 $\boldsymbol{DCN}^{(i,j)}$ 进行归一化处理，得到矩阵 $\boldsymbol{DCN}_{\boldsymbol{u}}^{(i,j)}$ ，其元素为 dcn'_{mn} ，具体的计算如公式(4)、(5)所示。

$$dcn'_{mn} = \frac{dcn_{mn}}{max_m} \tag{4}$$

$$max_m = max_n \left(dcn_{mn} \right) \tag{5}$$

其中 max_m 表示差值矩阵第 m 行的最大值。

定义规模为 l 的随机森林的相似性度量矩阵为 $\boldsymbol{R}_{\boldsymbol{F}}$ ， $\boldsymbol{R}_{\boldsymbol{F}}$ 的大小和树的数量有关，是 $l \times l$ 的方阵。其元素 rf_{ij} 与归一化差值矩阵 $\boldsymbol{DCN}_{\boldsymbol{u}}^{(i,j)}, i, j = 1, \cdots, l$ 的关系如公式(6)所示。

$$rf_{ij} = \begin{cases} 0, \ (i \geq j) \\ \left\| DCN'_{i,j} \right\|_F = \sqrt{\sum_m^M \sum_n^M dcn'^2_{mn}}, \ (i < j) \end{cases} \tag{6}$$

当 rf_{ij} 越小，则树 i 与树 j 的相似度越高，两个分类器对样本的分类结果越接近。

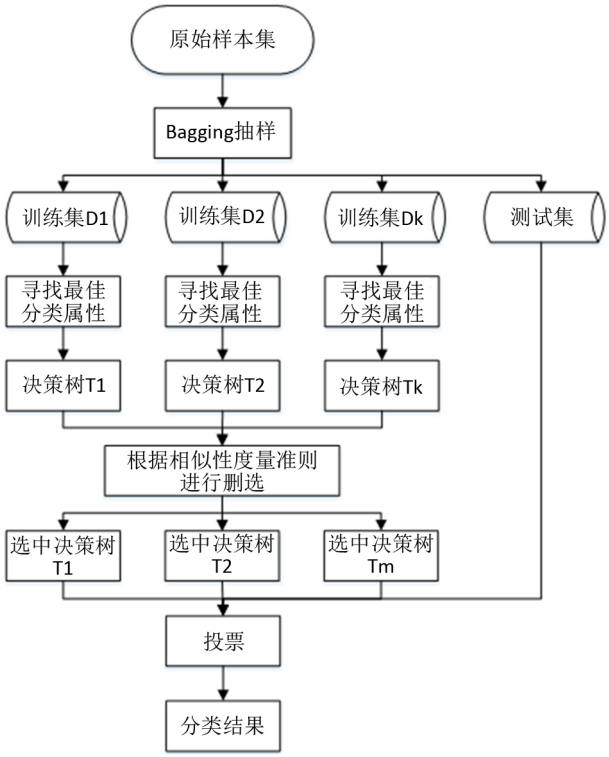

Figure 1. Random forest selection model
图 1. 随机森林选择模型

4.1.2. 基于“删劣”策略选择模型

对于常用选择策略，采用的是选优策略，即从分类器之中选择出若干代表性强的分类器。这种方法对分类器的分类效果和分类器之间的关联性都有要求，是多对多的关系，因此用该方法进行选择较为复杂[23]。本文选择另一种方法，即删除效果不佳的分类器。从基础分类器之中将相关度高的，分类能力差的分类器删除，然后将剩下的分类器集中在一起，组成新的模型。

这种删除策略只需要应用于相关度高的分类器之间，当相关度超过所设的阈值时，就将其剔除，因而该方法计算量将大大降低计算量。同时该方法还降低了总体分类器之间的相关度，从而提升分类能力。

4.1.3. 模型选择算法描述

具体步骤如算法 1 所示。

算法 1 随机森林模型选择算法

输入：决策树相似度阈值 t、分类准确度阈值 β

输出：随机森林模型 RF

1：通过决策树对测试样本进行分类预测；

2：根据分类结果，为决策树创建误差矩阵 $\boldsymbol{CN}$

3：创建相似度度量矩阵 $\boldsymbol{R}_F$：

4：for (($i, j = 1$ to l) & ($i < j$))

利用公式(3)计算 $\boldsymbol{DCN}^{(i,j)}$

利用公式(4)和公式(5)计算 $\boldsymbol{DCN}_u^{(i,j)}$

利用公式(6)计算 $\boldsymbol{DCN}_u^{(i,j)}$ 的范数，即 $\boldsymbol{R}_F$ 在该处的元素值

5：令 m_{ij} 为 $\boldsymbol{R}_F$ 中最小的非零元素

6：for ($m_{ij} < t$)

if (决策树 i 分类效果 $< \beta$)

将 $\boldsymbol{R}_F$ 中的树 i 清除

m_{ij} = $\boldsymbol{R}_F$ 中下一个最小非零元素

7：否则结束，未删除的决策树组成随机森林 RF。

5. 实验与分析

5.1. 实验数据说明

本文使用的数据集是 UCI 机器学习数据集的部分数据，具体信息如表 1 所示。

Table 1. Experimental data set taken from UCI
表 1. 取自 UCI 的实验数据集

UCI 数据集	属性类别	特征维数	样本个数	类别个数
Iris	连续	4	150	3
Breast-cancer	离散	10	286	2
Anneal	离散连续	38	798	8

5.2. 实验结果与分析

本实验通过从 10 到 100 内，不同的十个随机森林规模，然后对原始的随机森林(RF)和基于误差矩阵

的随机森林(CM-RF)的分类结果进行对比采用的指标为平均分类准确率。实验结果如图 2~4 示。

由实验结果看出，对于 iris 和 anneal 数据集，在初始树规模不同的情况下，基于误差矩阵的随机森林的平均分类准确率均高于传统的随机森林模型。对于 glass 数据集，随着随机森林在建数目的增加，传统的随机森林算法和本文提出的模型都出现了分类准确率下降的情况，但是，基于随机矩阵的森林的分类准确率下降更加缓慢，从而本文算法保持了一定的鲁棒性。因此，进一步说明本文提出的基于误差矩阵的随机森林模型的有效性。

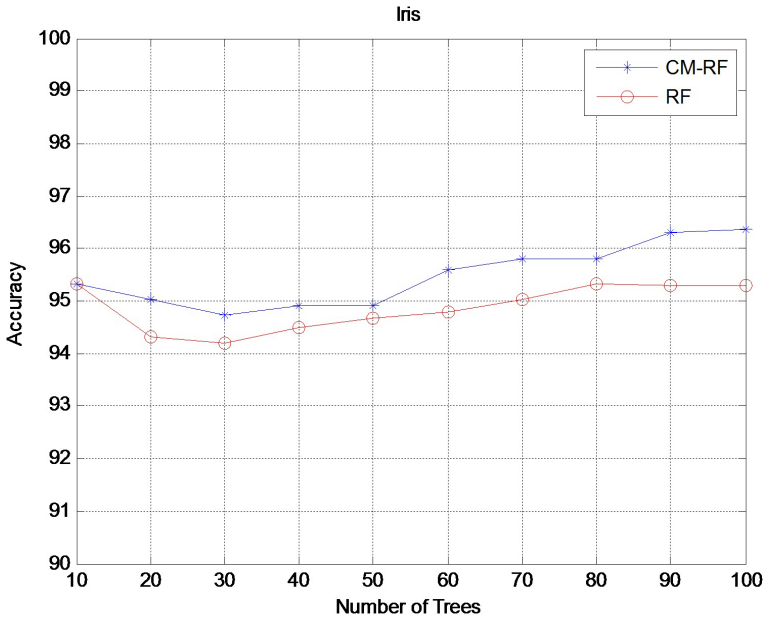

Figure 2. Accuracy comparison results on the Iris dataset
图 2. Iris 数据集上 Accuracy 对比结果

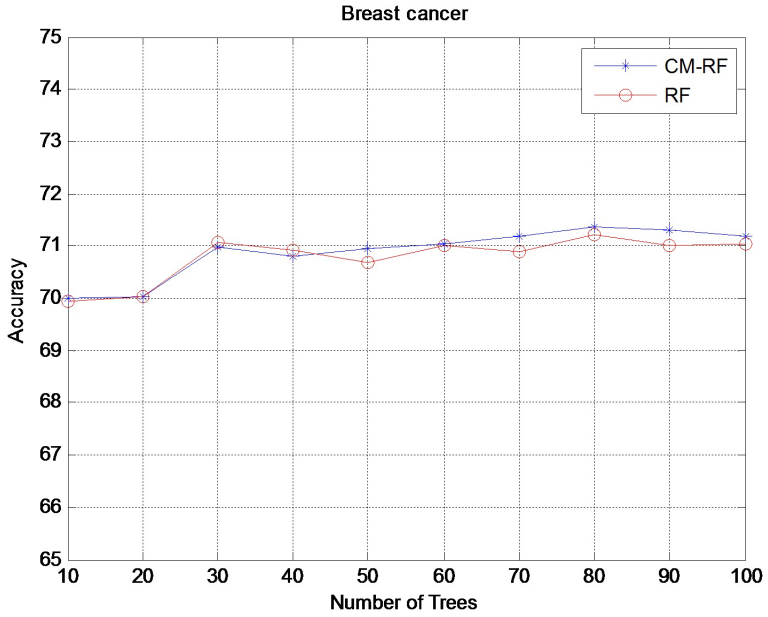

Figure 3. Accuracy comparison results on the Breast-cancer dataset
图 3. Breast-cancer 数据集上 Accuracy 对比结果

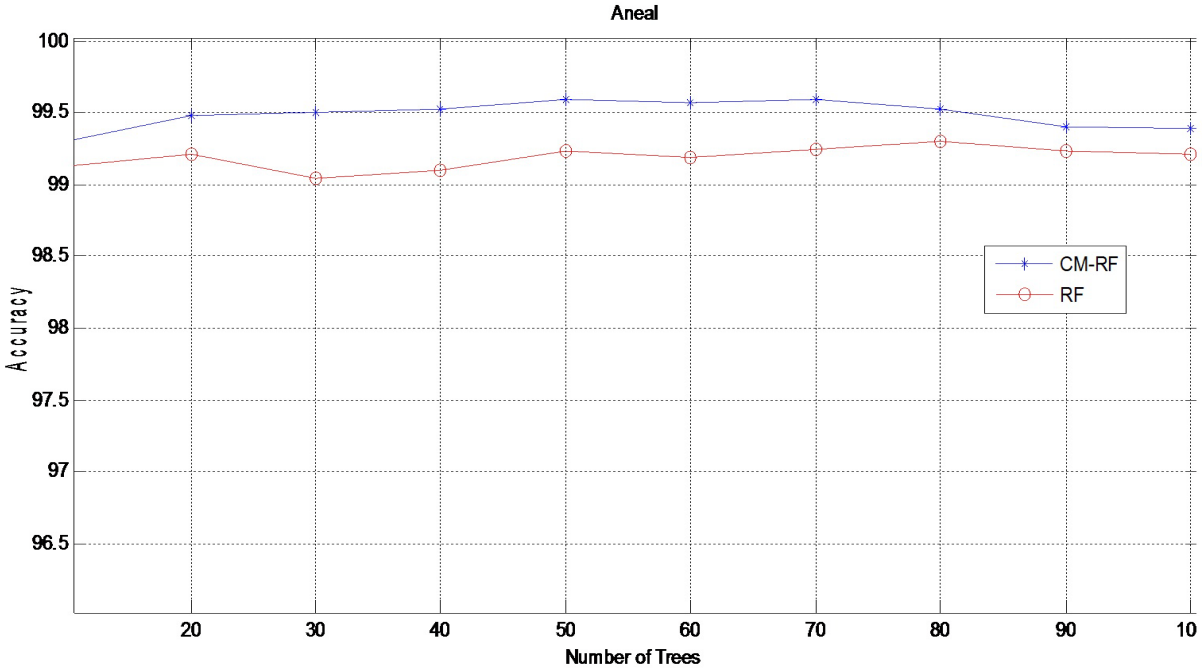

Figure 4. Accuracy comparison results on the Aneal dataset
图 4. Aneal 数据集上 Accuracy 对比结果

6. 结束语

本文提出了一种基于误差矩阵的随机森林分类模型选择方法。将误差矩阵应用于决策树的相似性度量中，通过使用矩阵的距离测度和向量夹角判断两棵树的相似性，考虑到树的数量占比问题，对矩阵每行进行归一化处理，之后结合决策树的分类性能，采用“删劣”思想完成随机森林模型的选择。由实验结果可知，该方法提高了分类的准确度。

基金项目

赛尔网络下一代互联网技术创新项目(NGII20170104)。

参考文献

[1] Babar, B., Luppino, L.T., Boström, T. and Anfinsen, S.N. (2020) Random Forest Regression for Improved Mapping of Solar Irradiance at High Latitudes. *Solar Energy*, **198**, 81-92. https://doi.org/10.1016/j.solener.2020.01.034

[2] Li, J., Tian, Y., Zhu, Y., Zhou, T.S., Li, J., Ding, K.F. and Li, J.S. (2020) A Multicenter Random Forest Model for Effective Prognosis Prediction in Collaborative Clinical Research Network. *Artificial Intelligence in Medicine*, **103**, Article ID: 101814. https://doi.org/10.1016/j.artmed.2020.101814

[3] Hammou, B.A., Lahcen, A.A. and Mouline, S. (2019) An Effective Distributed Predictive Model with Matrix Factorization and Random Forest for Big Data Recommendation Systems. *Expert Systems with Applications*, **137**, 253-265. https://doi.org/10.1016/j.eswa.2019.06.046

[4] Çifçi, M.A., Ertugrul, D.Ç. and Elçi, A. (2016) A Search Service for Food Consumption Mobile Applications via Hadoop and MapReduce Technology. 2016 *IEEE* 40*th Annual Computer Software and Applications Conference* (*COMPSAC*), Atlanta, 10-14 June 2016, 77-82. https://doi.org/10.1109/COMPSAC.2016.35

[5] 刘迎春, 陈梅玲. 流式大数据下随机森林方法及应用[J]. 西北工业大学学报, 2015, 33(6): 1055-1061.

[6] Yousfi, S. and Chiadmi, D. (2015) Big Data-as-a-Service Solution for Building Graph Social Networks. 2015 *International Conference on Cloud Technologies and Applications* (*CloudTech*), Marrakech, 2-4 June 2015, 1-6. https://doi.org/10.1109/CloudTech.2015.7337009

[7] 韩伟, 张学庆, 陈旸. 基于 MapReduce 的图像分类方法[J]. 计算机应用, 2014, 34(6): 1600-1603.

[8] Rajagopalan, M.R. and Vellaipandiyan, S. (2013) Big Data Framework for National E-Governance Plan. 2013 11*th International Conference on ICT and Knowledge Engineering* (*ICT&KE*), Bangkok, 20-22 November 2013, 1-5. https://doi.org/10.1109/ICTKE.2013.6756283

[9] 孙悦, 袁健. 基于 Spark 的改进随机森林算法[J]. 电子科技, 2019, 32(4): 60-63+67.

[10] Gall, J., Yao, A., Razavi, N., Cool, L.V. and Lempitsky, V. (2011) Hough Forests for Object Detection, Tracking, and Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, 2188-2202. https://doi.org/10.1109/TPAMI.2011.70

[11] Gall, J. and Lempitsky, V. (2009) Class-Specific Hough Forests for Object Detection. 2009 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 1022-1029. https://doi.org/10.1109/CVPR.2009.5206740

[12] Ishwaran, H., Kogalur, U.B., Xi C. and Minn, A.J. (2011) Random Survival Forests for High-Dimensional Data. *Statistical Analysis and Data Mining*, **4**, 115-132. https://doi.org/10.1002/sam.10103

[13] 谢晓龙, 叶笑冬, 董亚明. 梯度提升随机森林模型及其在日前出清电价预测中的应用[J]. 计算机应用与软件, 2018, 35(9): 327-333.

[14] 魏正韬. 基于非平衡数据的随机森林算法研究[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2017.

[15] 王诚, 高蕊. 基于特征约减的随机森林改进算法研究[J/OL]. 计算机技术展, 2020, 30(3): 40-45.

[16] 雍凯. 随机森林的特征选择和模型优化算法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2008.

[17] 毕凯, 王晓丹, 姚旭, 等. 一种基于 Bagging 和混淆矩阵的自适应选择性集成[J]. 电子学报, 2014(4): 711-716.

[18] Tumer, K. and Ghosh, J. (1996) Error Correlation and Error Reduction in Ensemble Classifiers. *Connection Science*, **8**, 385-340. https://doi.org/10.1080/095400996116839

[19] Faria, F.A., Dos Santos, J.A., Sarkar, S., *et al*. (2013) Classifier Selection Based on the Correlation of Diversity Measures: When Fewer Is More. 2013 *XXVI Conference on Graphics, Patterns and Images*, Arequipa, 5-8 August 2013, 16-23. https://doi.org/10.1109/SIBGRAPI.2013.12

[20] Shi, H.L., Ferguson, D., Beagley, J. and Huyck, M. (2008) Work in Progress—Improving Interrater Agreement Used to Measure Learning Outcomes. 2008 38*th Annual Frontiers in Education Conference*, Saratoga Springs, 22-25 October 2008, F2B-7-F2B-8. https://doi.org/10.1109/FIE.2008.4720398

[21] Löfström, T., Johansson, U. and Boström, H. (2008) On the Use of Accuracy and Diversity Measures for Evaluating and Selecting Ensembles of Classifiers. 2008 7*th International Conference on Machine Learning and Applications* (*ICLMLA*), San Diego, 11-13 December 2008, 127-132. https://doi.org/10.1109/ICMLA.2008.102

[22] 乔少杰, 唐常杰, 陈瑜, 等. 基于树编辑距离的层次聚类算法[J]. 计算机科学与探索, 2007, 1(3): 282-292.

[23] 谢元澄, 杨静宇. 删除最差基学习器来层次修剪 Bagging 集成[J]. 计算机研究与发展, 2009, 46(2): 261-267.