计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(9):193−198 [doi: 10.15888/j.cnki.csa.006537]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进随机森林算法的图像分类应用①

张志禹, 吉元元, 满蔚仕

(西安理工大学 自动化与信息工程学院, 西安 710048)

摘　要: 针对随机森林算法中节点分裂方式单一且相似的问题, 提出一种改进节点分裂方式的优化算法, 将算法中独立的节点分裂方式 ID3 与 CART 进行重新组合, 通过自适应参数选择得到新的分裂规则, 用于最优属性的选择划分并应用于图像分类问题. 首先以词袋模型为基础, 加入空间金字塔结构来提取图像特征, 并将其量化成视觉词汇, 最后结合 Spark 平台用改进节点分裂方式的随机森林算法实现图像分类. 实验结果表明, 通过选择组合算法的最优系数, 该算法有效提高图像分类准确率, 并保证算法运行效率.

关键词: 图像分类; 随机森林; 节点分裂; 空间金字塔

引用格式: 张志禹,吉元元,满蔚仕.改进随机森林算法的图像分类应用.计算机系统应用,2018,27(9):193–198. http://www.c-s-a.org.cn/1003-3254/6537.html

Image Classification Application Based on Improved Random Forest Algorithm

ZHANG Zhi-Yu, JI Yuan-Yuan, MAN Wei-Shi

(Faculty of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: An improved random forest node splitting algorithm is proposed in this study for improving the accuracy of image classification. The independent splitting method ID3 and CART are re-combined, and new splitting rules are obtained by adaptive parameter selection. On the basis of the bag-of-words model, the spatial pyramid model is introduced to extract image features. After dividing the image into different grids, k-means algorithm is then used to character clustering. Finally, it uses the algorithm for verification on a large number of images on Spark. The results show that the algorithm can be applied to distributed systems, and can greatly improve the classification accuracy while ensuring the efficiency of the algorithm at the same time.

Key words: image classification; random forest; node splitting; spatial pyramid model

1　引言

随着互联网技术、多媒体应用和计算机视觉的不断发展, 对于海量场景图像的分类处理成为不容小觑的问题. 近年来, 主要以词袋模型 (Bag of Word, BoW)、卷积神经网络等图像分类算法的有效分类性能吸引了更多的关注. 图像分类已成为管理应用图像数据的关键技术, 由于图像的多样性和复杂性以及类内的差异性, 如何更加准确全面地表示图像是一个问题. 早期的图像分类是通过提取图像的底层特征, 如颜色、纹理等特征. 但是, 这些算法对应的是全局信息从而确定目标的整体结构不能变, 且会因为图像缺失或者光线或遮挡问题而受到影响, 这样在处理复杂图像时效果并不理想. Avila[1]在图像分类中用到了词袋模型, 并且引入了基于密度函数的池策略. 这种方法能够更好地代表词典的码字并描述图像. 将该方法用在视频和图像分类上, 都有不错的分类效果. Li 等人[2]将视觉词汇与

① 基金项目: 国家自然科学基金 (41390454)
Foundation item: National Natural Science Foundation of China (41390454)
收稿时间: 2018-01-23; 修改时间: 2018-02-27; 采用时间: 2018-03-06; csa 在线出版时间: 2018-08-16

空间金字塔匹配模型结合, 提出了一种仿射传播聚类算法用于高分辨率遥感图像分类, 实验结果表明该算法分类性能优于传统聚类算法.

随机森林算法在处理非平衡数据集、连续变量与决策树节点分裂算法[3]问题等方面提出和发展了许多新方法. 对场景图像进行特征提取后的后续分类, 本文拟采用随机森林 (Random Forest, RF) 算法做进一步的研究. 文献[4]中提出一种新的特征加权方法和决策树选择方法 (Improved Random Forest, IMRF), 结合协同服务, 使随机森林算法适用于多类大量图像数据的分类. 利用该方法, 在不增加误差界的前提下, 有效地减少子空间的大小, 提高分类性能. Archana Chaudhary 等人[5]由随机森林机器学习算法、属性评估方法和实例过滤方法组成一种新的随机森林分类器方法, 并用于多类别花生病害分类问题, 并极大提高分类精度. 但是, 这些方法在海量数据的分类效率与分布式计算问题上还存在一定的制约, 同时分类精度也有待进一步提高, 难以适应信息量的爆炸式增长, 因此相关问题上还有待进一步学习研究.

Apache Spark 集群计算平台[6,7](如图 1) 是一个基于内存计算的开源运算系统, 在运算速度上可以满足人们的需要; Spark 启用了内存分布数据集[8], 除了能够提供交互式查询外, 它还可以优化迭代工作负载,具有很好的容错机制[9], 该机制可以维护 “血统”,可以记录特定数据转换操作行为的过程. 同时 Spark 可以很好的兼容 Hadoop 生态系统, 这使得其应用发展都有了很好的基础. 因此本文中, 有关于场景图像分类的若干步骤将在该平台下进行, 有利于对大数据量问题的研究与分布式计算的实现.

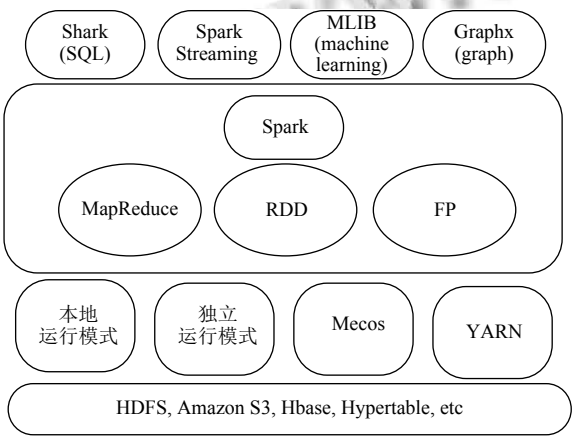


图 1 Spark 生态系统

在本文中实现图像分类的步骤如下:

Step1. 利用 SURF 特征进行图像特征采样[10], 再利用局部特征描述子形成对这些向量的表达;

Step2. 对图像的特征向量进行聚类得到视觉单词[11], 计算每幅图片到这些视觉单词的距离, 并将其映射到距离最近的视觉单词, 完成每幅图像的词频表达[12];

Step3. 利用改进的自适应节点分裂随机森林算法 (Self-Adaptive Node Split Random Forest, SANS-RF) 进行图像分类并利用包外图像进行验证, 改进算法及涉及到的理论会在后续段落重点介绍.

2 空间金字塔模型

2.1 词袋模型

在场景图像分类的众多算法中, BoW 模型的最大优点是将图像表示为视觉词汇, 更容易识别并表示出图像中感兴趣的部分[13], 即将图像看作一个“文档”, 关键词就是提取图像的 SURF 特征, 称为“视觉词典”[12].

为了在特征点检测与匹配实现尺度不变性, SURF 算法首先用 Hessian 矩阵确定候选点, 然后进行非极大抑制, 会使计算复杂度降低许多. Hessian 矩阵是 SURF 算法的核心, 即根据图像中每一个像素点的 Hessian 矩阵, 如式 (1), 得到 Hessian 判别式, 如式 (2), 其值即是 Hessian 矩阵的特征值, 可以用该式的结果对像素点进行分类:

$$\mathrm{H}(f(x,y)) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x \partial y} \\ \dfrac{\partial^2 f}{\partial x \partial y} & \dfrac{\partial^2 f}{\partial y^2} \end{bmatrix} \tag{1}$$

$$\det(\mathrm{H}) = \frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 \tag{2}$$

在 SURF 算法中, 通常利用图像像素$I(x,y)$代替原始的$f(x,y)$, 通过特定核间的卷积计算二阶偏导数, 可以得到 Hessian 矩阵的三个元素L_{xx},L_{yy},L_{xy}, 因此 Hessian 矩阵如下所示:

$$\mathrm{H}(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \tag{3}$$

同时选用二阶标准高斯函数作为滤波器, 即在 Hessian 矩阵构造前, 需对其进行高斯滤波:

$$L(x,t) = G(t) \cdot I(x,t) \tag{4}$$

其中 $L(x,t)$ 代表一幅图像在不同解析度下的表示, $G(t)$ 代表高斯核, 公式如下:

$$G(t)=\frac{\partial^2 g(t)}{\partial x^2} \tag{5}$$

以上计算可以判别特征点, 为此 Herbert Bay[14]提出用近似值代替 $L(x,t)$, 为减小准确值与近似值之间的误差引入权值,权值随尺度变化, 则 Hessian 矩阵的判别式表示为:

$$det(\mathrm{H}_{\mathrm{approx}})=D_{xx}D_{yy}-(0.9D_{xy})^2 \tag{6}$$

具体公式推导可详见文献[14].

通过以上方法可以生成尺度空间, 再通过精确定位特征点, 选取特征点主方向确定的步骤, 就可以构造 SURF 特征点描述算子, 进行图像特征提取.

2.2 空间金字塔结构

利用上一小节提到的词袋模型表示图像可以得到一个不错的分类效果,但是该模型没有考虑图像的空间位置信息, 得到的是图像的一个无序集合. 因此在这一步骤中引入了空间金字塔模型, 以达到充分利用图像空间信息的要求.

该模型首先对局部特征量化, 然后在每个金字塔水平把图像划分为细网格序列[15], 从每个金字塔水平的网格中提取特征, 同时给每层网格分配一个权重, 按权重把每层网格特征加权串联在一起, 如图 2 所示.

图 2 空间金字塔模型示意图

F 表示金字塔分割下图像的特征向量:

$$F=\left(f_1^0,f_1^1,\cdots,f_{c(1)}^1,\cdots,f_1^{L-1},\cdots,f_{c(L-1)}^{L-1}\right) \tag{7}$$

W_l 表示不同金字塔层次的加权值:

$$W_l=\begin{cases}\dfrac{1}{2^L}, l=0\\ \dfrac{1}{2^{L-l+1}}, l\neq 0\end{cases} \tag{8}$$

所以一幅图像的最终加权空间金字塔表现方法为:

$$f_W{}^{(N_\omega)}=FW_l=((F\omega_1)^{\mathrm{T}},\cdots,(F\omega_{N_\omega})^{\mathrm{T}})^{\mathrm{T}}, f_W{}^{N_W}\in R^{N_W d} \tag{9}$$

以上公式可以将需要分类的图像更好表示.

3 随机森林算法

3.1 算法简介

随机森林是一种组合分类器, 它利用 Boostrap 重抽样方法从原始样本中抽取多个样本[16]构造子数据集, 利用子数据集形成基决策树并对其进行训练, RF 在决策树的训练中引入了随机属性选择, 即对基决策树的每个节点, 先从该节点的属性集合中随机选择一个包含 k 个属性的子集, 然后再从这些子集中选择一个最优属性用于节点分裂, 这样可以使每棵决策树彼此不同, 提升系统的多样性, 然后将这些决策树组合在一起, 利用 Boostrap 中未抽取到的样本作为包外数据集进行验证, 并通过投票法得到分类结果, 从而提升分类性能, 算法流程图如图 3 所示.

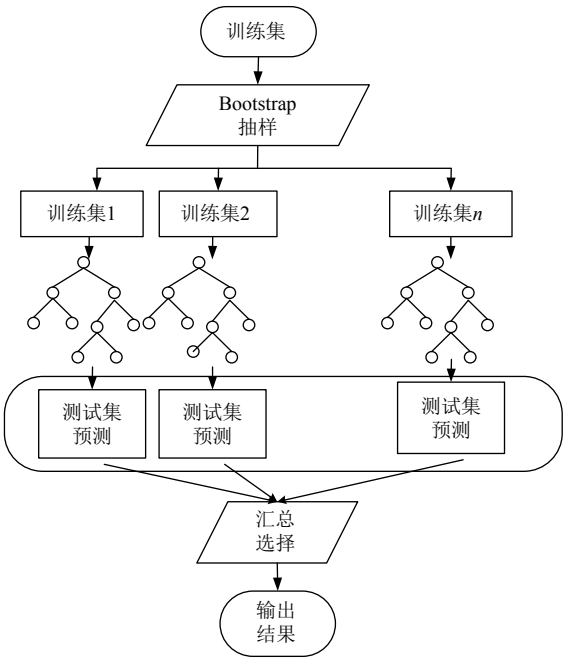


图 3 随机森林算法

节点分裂是 RF 算法的核心步骤, 通过节点分裂才能产生一颗完整的决策树[17]. 每棵树分支的生成, 都是按照某种分裂规则选择属性, 这些规则主要包括信息增益最大、信息增益率最大和 Gini 指数最小等原则, 然后选择某个属性作为分裂属性, 并按照其划分实现决策树分支生长. 随着划分过程的进行, 节点的纯度越来越高, 即该节点所包含的样本尽可能的属于同一

类别.

3.2 改进随机森林算法

大量研究都证明了随机森林算法具有较高的分类准确率, 对异常值和噪声有很好的容忍度, 而且不易出现过拟合. 本文提出的 SANS-RF 算法, 通过参数的自适应选择过程, 来优化算法中决策树的节点分裂算法, 达到提高算法分类精度的目的.

对同一个数据集, 选择不同的节点分裂算法, 也会因选择的属性不相同而得到不同的决策树, 得出随机森林的分类精度会有差异. 因此提出在生成决策树时, 选择最优的属性进行节点分裂, 即将节点分裂算法进行线性组合, 形成新的分裂规则, 应用于节点属性的选择划分. 由于 Spark mllib 的随机森林算法中集成的节点分裂算法只有 ID3 和 CART, 因此节点分裂优化的考虑暂定这两种算法上, 其节点分裂公式表示用属性a对样本集D进行划分所获得的信息增益与基尼指数分别如下:

$$Gain(D,a)=Ent(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}Ent(D^v) \tag{10}$$

$$Gini(D,a)=\sum_{v=1}^{V}\frac{|D^v|}{D}Gini(D^v) \tag{11}$$

其中D^v表示第v个分支节点包含的D中所有在属性a上取值为a^v的样本:

$$Ent(D)=-\sum_{k=1}^{|y|}p_k\log_2 p_k \tag{12}$$

$$Gini(D)=\sum_{k=1}^{|y|}\sum_{k'\neq k}p_k p'_k=1-\sum_{k=1}^{|y|}p_k^2 \tag{13}$$

式 (12) 和式 (13) 分别表示数据集 D 的信息熵与基尼值.

表 1 节点分裂算法对比

算法	节点分裂准则	准则指标
ID3	信息增益最大	划分的数据集中样本的纯度
CART	Gini 指数最小	从数据集中随机抽两个样本不同的概率

结合表 1 内容, 节点分裂准则应以划分后数据集纯度更高为目标, 因此组合节点分裂公式为:

$$H=\min_{\alpha,\beta\in R}F\{D,a\}=\alpha Gini(D,a)-\beta Gain(D,a)$$
$$\text{s.t.}\begin{cases}\alpha+\beta=1\\0\le\alpha,\beta\le 1\end{cases} \tag{14}$$

其中, 参数α,β代表两种算法在$H(x)$中的系数, 式中取H值为最小, 即 ID3 与 CART 均最优作为节点划分准则可提升分类效果.

由于不同图像集中图像的特征是不同的, 所以 SANS-RF 算法中的参数选择也难以固定, 因此采用自适应参数选择过程, 得出最优的组合参数, 对于参数α,β应满足上式中的约束条件.

实验中采用分类错误率与准确率进行性能度量, 对于样本D, 分类错误率定义为:

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{II}(f(x_i)\neq y_i) \tag{15}$$

准确率则定义为:

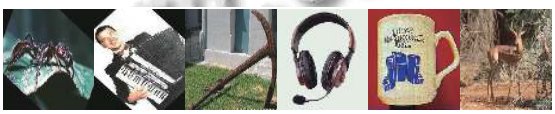
$$acc(f;D)\frac{1}{m}\sum_{i=1}^{m}\mathrm{II}(f(x_i)=y_i)=1-E(f;D) \tag{16}$$

具体实验效果在下节进行对比验证.

4 实验过程及结果

4.1 空间金字塔模型

本节通过对比实验来验证词袋模型与空间金字塔模型的分类效果, 实验设置为对 Caltech101, 256_ObjectCategories, SUN2012 三种数据集中如图 4 所示, 对这些图像提取特征并聚类, 最后利用包外数据进行测试得到分类错误率 testErr, 每组实验进行多次取平均值作为最终实验数据, 实验结果如图 5 所示.



(a) Caltech 101 数据集

(b) 256_Object Categories 数据集

(c) SUN 2012 数据集

图 4 数据集样本

从图 5 中数据可以看出对这三种数据集, 在词袋模型的基础上引入空间金字塔模型可以有效的提高分

类准确度, 降低错误率, 因此在后续算法改进中会以此模型为基础继续进行.

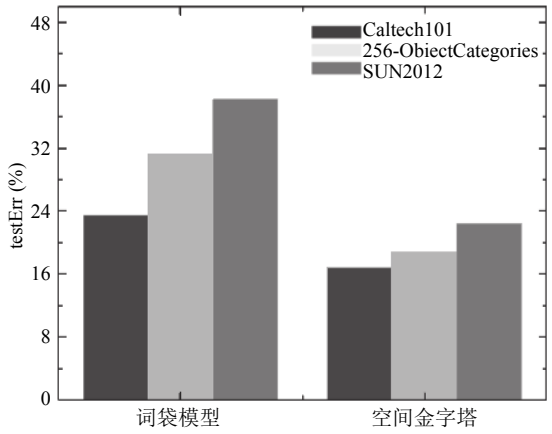


图 5 空间金字塔与词袋模型对比结果

4.2 分布式 vs 单机版

图像分类算法的计算时间会随着图片数量增加而急剧增加, 但是在大数据平台下, 可以利用分布式处理来缩短程序的运行时间, 该平台有三个节点分别为 master, slave1, slave2, 其内存为 8 GB, 4 线程运行, 同时将图片的视觉特征文件存放在 Hadoop HDFS 分布式系统中, Spark 单机版与分布式系统运行对比结果见表 2, 运行时间以分钟为单位.

表 2 单机与分布式运行时间对比

图像数	200	500	750	1000
单机	35	61	85	120
分布式	16	23	30	41

加速比是指同一个任务在单机系统和分布式系统中运行所用时间的比率, 用来衡量分布式算法的效率, 其计算公式为 $Sp=T1/T2$, $T1$ 是单节点下运行时间, $T2$ 是分布式运行时间, 结果如图 6 所示.

4.3 改进随机森林算法的结果

根据上一节中 SANS-RF 算法的改进公式可知, 线性组合算法的系数值对分类结果会有重要的影响, 因此本节中首先用不同图像集中的 1000 幅图片进行测试, 人为给定参数值, 并以包外数据的分类错误率 testErr 作为指标进行验证, 实验结果如表 3 所示.

由表 3 可知对不同图像集参数的最优组合是不能固定的, 因此引入参数的自适应选择来得到最优的分类结果是合理的.

SANS-RF 算法的在三种不同图像集上的分类结果如图 7 至图 9 所示, 其中, SVM (Support Vector Machine) 是通常情况下图像分类会选择的算法, 原始 RF 指 Spark 平台上未改进的随机森林方法, IMRF 为文献[4]中提出的利用权重与决策树选择的随机森林改进算法.

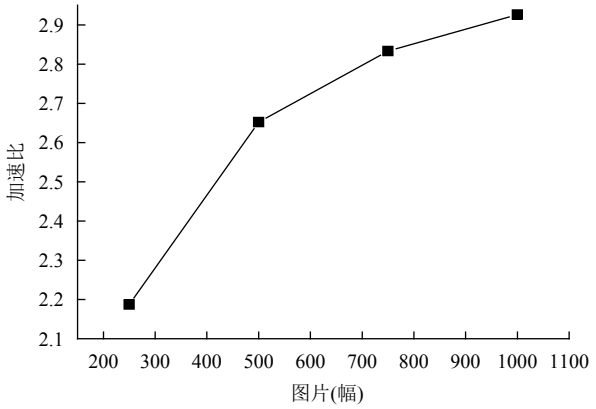


图 6 Spark 平台加速比结果图

表 3 SANS-RF 算法参数验证表

α	β	256_ObjectCategories	Caltech101	SUN2012
0.0	1.0	5.201	12.903	16.053
0.2	0.8	6.2292	6.035	10.968
0.4	0.6	3.431	2.624	5.214
0.6	0.4	6.248	7.462	9.431
0.8	0.2	12.691	13.251	21.638
1.0	0.0	18.871	18.533	29.181

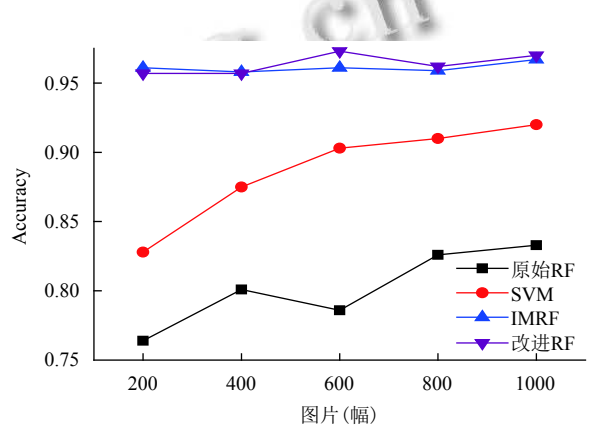


图 7 图像集 1(Caltech-101) 中算法分类准确率对比

通过这几种算法的对比, 实验结果表明, 本文中提出的 SANS-RF 算法有着很好的分类准确率,远远高于基础 RF 算法与支持向量机分类效果 , 并且比 IMRF 算法更加稳定, 更适用于海量图像的分布式应用. 因此, 本文提出的基于 Spark mllib 随机森林的组合节点分裂算法是令人满意的.

5 结束语

本文在 Spark 平台下实现了不同场景图像的准确分类, 首先在简单的词袋模型的基础上验证了空间金字塔模型的有效性; 其次针对随机森林的节点分裂算法进行改进并实验, 通过对比, 验证该算法的有效性与准确性. Spark 平台可以有效提高算法运行效率的同时, 又保证了分类准确率, 适合海量图像的分类研究.

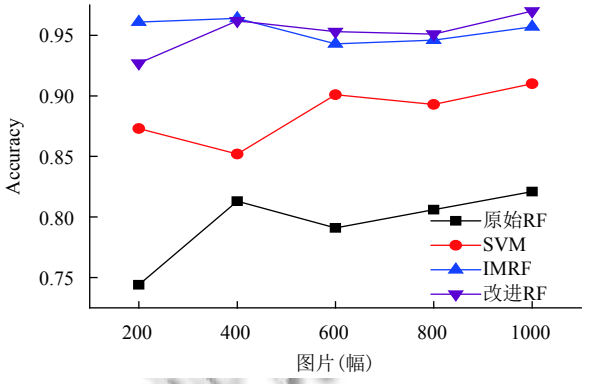


图 8 图像集 2(256-ObjectCategories) 中算法分类准确率对比

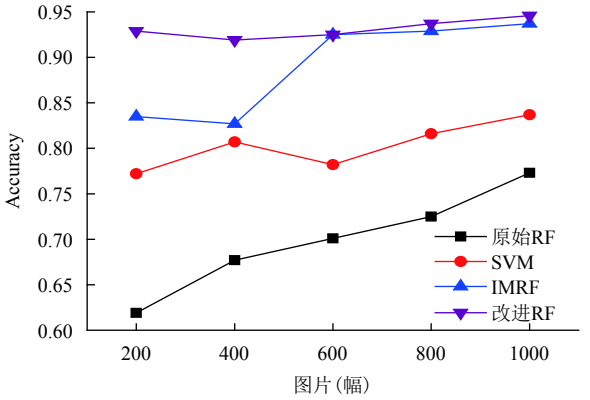


图 9 图像集 3(SUN2012) 中算法分类准确率对比

同时可以在增加分类图片数量和融合更成熟有效的节点分裂算法上进一步研究, 以体现 Spark 平台在处理速度上的优势, 并提高分类准确率.

参考文献

1 Avila S, Thome N, Cord M, *et al*. BOSSA: Extended bow formalism for image classification. 2011 18th IEEE International Conference on Image Processing. Brussels. 2011. 2909–2912. [doi:10.1109/ICIP.2011.6116268]

2 Li X, Zhang L, Wang L, *et al*. Effects of BOW model with affinity propagation and spatial pyramid matching on polarimetric SAR image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2017, 10(7): 3314–3322. [doi: 10.1109/JSTARS.2017.2671364]

3 李慧, 李正, 佘堃. 一种基于综合不放回抽样的随机森林算法改进. 计算机工程与科学, 2015, (7): 1233–1239. [doi: 10.3969/j.issn.1007-130X.2015.07.002]

4 Xu B, Ye Y, Nie L. An improved random forest classifier for imageclassification. International Conference on Information and Automation. IEEE. 2012. 795–800. [doi: 10.1109/ICInfA.2012.6246927]

5 Chaudhary A, Kolhe S, Kamal R. An improved Random Forest Classifier for multi-class classification. Information Processing in Agriculture, 2016, 3(4): 215–222. [doi: 10.1016/j.inpa.2016.08.002]

6 Reyes-Ortiz JL, Oneto L, Anguita D. Big data analytics in the cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf. Procedia Computer Science, 2016, 53: 121–130.

7 Singh S, Liu Y. A cloud service architecture for analyzing big monitoring data. Tsinghua Science and Technology, 2016, 21(1): 55–70. [doi: 10.1109/TST.2016.7399283]

8 Islam NS, Wasi-Ur-Rahman M, Lu X, *et al*. Performance Characterization and acceleration of in-memory file systems for hadoop and spark applications on HPC clusters. IEEE International Conference on Big Data. 2015. 243–253. [doi: 10.1109/BigData.2015.7363761]

9 Yigitbasi N, Willke TL, Liao GD, *et al*. Towards machine learning-based auto-tuning of MapReduce. IEEE 21st International Symposium on Modelling. 2013. 11–20. [doi: 10.1109/MASCOTS.2013.9]

10 朱杰, 超木日力格, 谢博鋆, 等. 利用颜色进行层次模式挖掘的图像分类方法. 计算机科学与探索, 2017, (3): 396–406.

11 Kausar N, Majid A, Javed SG. Developing multi-focus image fusion system with random forest learning algorithm for real-blurred images. 2016 13th International Bhurban Conference on Applied Sciences and Technology. 2016. 219–224. [doi: 10.1109/IBCAST.2016.7429880]

12 Kurinjivendhan N, Thangadurai K. Modified k-means algorithm and genetic approach for cluster optimization. 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). 2016. 53–56.

13 Gaitán D, Isaza C, Gómez W, *et al*. Categorization of ecosystems based on soundscape analysis: A perspective from image classification. International Conference on Computational Science and Computational Intelligence. IEEE. 2017. 762–766.

14 The offical home of the image processing libary. http://www.chrisevansdev.com/computer-vision-opensurf.html.

15 Abdelouahed S, Ennouni A, Aarab A. Automatic estimation of clusters number for K-means. 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt). 2016. 450–454. [doi: 10.1109/CIST.2016.7805089]

16 郭佳. 场景图像分类的相关技术研究[硕士学位论文]. 西安: 西安电子科技大学, 2013.

17 曹正凤. 随机森林算法优化研究[博士学位论文]. 北京: 首都经济贸易大学, 2014.