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.

ant accordion spear earphone cup gerenuk

(a) Caltech 101 数据集

bear breadmaker eyeglasses mailbox pyramid treadmill

(b) 256_Object Categories 数据集

abbey doorway jetty towel river sauna

(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.

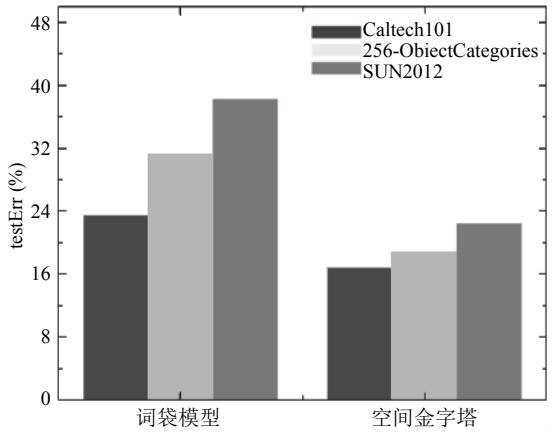

图 5 空间金字塔与词袋模型对比结果

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.

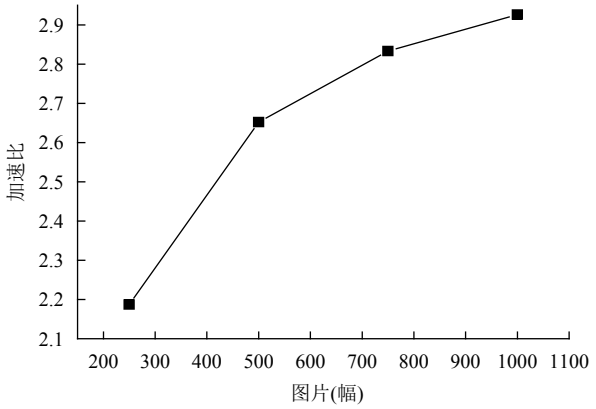

图 6 Spark 平台加速比结果图

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

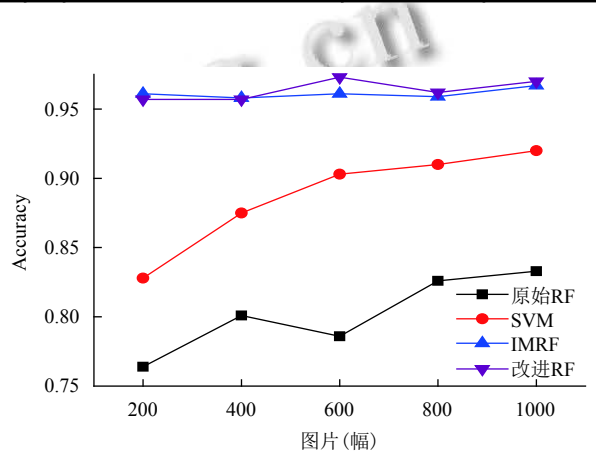

图 7 图像集 1(Caltech-101) 中算法分类准确率对比

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.

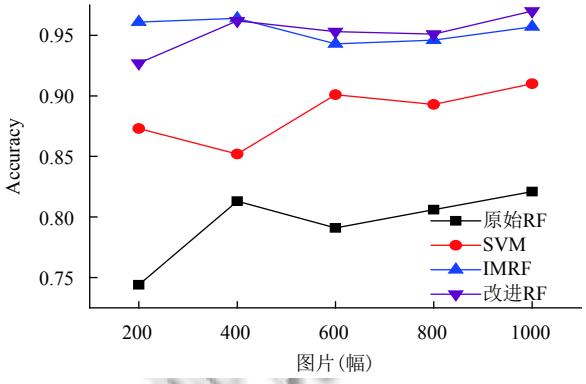

图 8 图像集 2(256-ObjectCategories) 中算法分类准确率对比

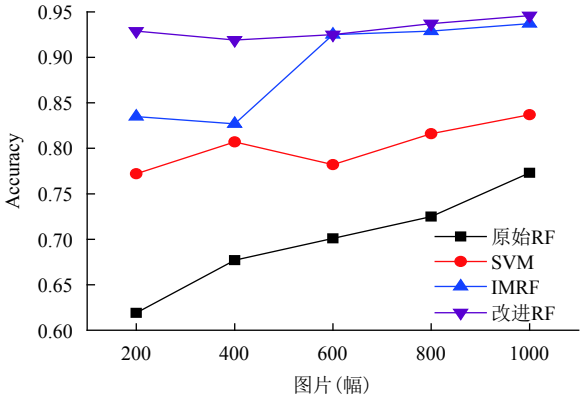

图 9 图像集 3(SUN2012) 中算法分类准确率对比

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.