基于弱相关化特征子空间选择的离散化随机森林并行分类算法

陈旻骋 袁景凌 王啸岩 朱 赛
(武汉理工大学计算机科学与技术学院 武汉 430070)

摘 要 随着大数据时代的到来,数据信息呈几何倍数增长。传统的分类算法将面临着极大的挑战。为了提高分类算法的效率,提出了一种基于弱相关化特征子空间选择的离散化随机森林并行分类算法。该算法在数据预处理阶段对数据集中的连续属性进行离散化。在随机森林抽取特征子空间阶段,利用属性向量空间模型计算属性间的相关性,构造弱相关化特征子空间,使所构建的决策树之间相关性降低,从而提高随机森林的分类效果;并通过研究随机森林的并行化策略,结合MapReduce框架,改进并实现了随机森林模型构建过程的双重并行化,进一步改善了算法的计算效率。

关键词 随机森林,离散化,弱相关化特征子空间,并行分类
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.011

Parallelization of Random Forest Algorithm Based on Discretization and Selection of Weak-correlation Feature Subspaces

CHEN Min-cheng YUAN Jing-ling WANG Xiao-yan ZHU Sai
(School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China)

Abstract With the coming of the big data age, data information is increasing exponentially at a dramatic rate. The traditional classification algorithm will encounter great challenges. In order to improve the efficiency of classification algorithm, this paper proposd a parallel random forest algorithm based on discretization and the selection of the weak-correlation feature subspaces. This algorithm discretizes continuous attributes in data pretreatment phase. At the step of the selection of feature subspaces for growing decision trees, we used vector space modal of attributes to calculate the correlation between attributes, and then constructed the weak-correlation feature subspaces. This algorithm not only reduces the correlation among decision trees, but also improves the classifying effect of the random forest. We also designed and realized a double parallel method for building random forest model based on the MapReduce framework. This strategy goes a step further with its own charity efforts.

Keywords Random forest, Discretization, Weak-correlation feature subspaces, Parallel classification

1 引言

随着科学技术的不断发展,大数据已然成为如今数字化时代的重要标志之一。如何高效地分析和处理这些数据显得尤为重要。分类算法作为数据挖掘的核心技术,传统的分类算法,如决策树算法、支持向量机算法、贝叶斯分类算法等,在针对低维的小数据集时能获得较为满意的效果,但是当数据集的规模增大、数据的维数增高、数据的结构变得复杂时,传统算法的性能便会显著地降低[1]。

Leo Bremen于2001年提出的随机森林算法[2]由于其训练算法简单,预测算法的速度快,对噪音有着较好的耐受能力且对高维数据分类问题具有良好的可扩展性,因此在实际生活中得到了广泛的应用,成为数据分析、知识管理、模式识别等众多领域的研究人员和技术人员共同关注的一个热点研究话题。例如,在生物信息学方面,文献[3]将随机森林算法应用于多元神经影像表型的全基因组关联分析;在生态学方面,文献[4]利用随机森林算法对云南松分布进行模拟预测;在医学方面,文献[5]将随机森林算法应用于对肺结节的自动检测;在社会网络学方面,文献[6]将随机森林算法用于微博转发预测;在电气工程领域,文献[7]将随机森林算法用于电力用户对大数据的负荷预测。同时随机森林算法也经常运用于视觉处理的具体问题中。计算机视觉领域的顶级国际会议ICCV(International Conference on Computer Vision)在2013年收录了3~4篇使用此算法的文章。此外,随机森林在入侵检测[8,9]、故障诊断[10]方面也有着广泛的应用。但在国内,目前对随机森林的有关研究还处于起步阶段,因此对随机森林

到稿日期:2015-07-13 返修日期:2015-09-01 本文受国家自然科学基金(61303029),湖北省自然科学基金(2014CFB836),教育部留学回国人员科研启动基金([2012]1707)资助。

陈旻骋(1990-),男,硕士生,CCF会员,主要研究方向为数据挖掘、绿色计算,E-mail:wester589@gmail.com;袁景凌(1975-),女,博士,教授,博士生导师,CCF高级会员,主要研究方向为机器学习、大数据挖掘、绿色计算,E-mail:yuanjingling@126.com(通信作者);王啸岩(1991-),男,硕士生,主要研究方向为绿色计算、机器学习;朱 赛(1990-),女,硕士生,主要研究方向为数据挖掘、绿色计算。

的理论及其应用的深入探索变得尤为重要。

本文在传统随机森林算法的基础上，提出了一种基于弱相关化特征子空间选择的离散化随机森林并行分类算法（简称 DWRF 算法），主要工作如下：

（1）在数据预处理阶段对数据集中的连续属性离散化，避免了构建森林时连续属性的多次分裂，从而减少了森林的空间开销。

（2）结合统计学方法中的向量空间模型，提出了属性向量空间模型，计算属性间的相关性；并将其应用于随机森林抽取特征子空间阶段，构造弱相关化特征子空间，提高了森林的分类效果。

（3）设计及实现了基于 MapReduce 的并行化 DWRF 算法，并提出了双重并行化的建树方法，进一步提高了算法的运行效率。

2 算法改进策略

2.1 改进策略 1：离散化连续属性

在使用随机森林分类过程中，连续属性值的划分是二元划分的，根据分裂点的值将数据元组分为大于分裂值和小于分裂值的两个子集，采用这种方法建树的开销相对较高，且产生的决策树的分类效率不高。因此，需要对数据集进行预处理，如果数据集中包含连续属性，需将其离散化。Fayyad 等人证明：无论用于学习的数据集包含多少类别，类别怎样分布，一般情况下，连续属性的最佳分裂点在边界点处[11]。因此根据 Fayyad 边界点原理将数据集中的连续属性进行排序，随机抽取排序后某一连续型属性的相邻两类边界区$[a, a_n]$处的 k 个连续属性值$\{a_1, a_2, \cdots, a_k\}$作为测试属性组（$k$ 根据实际数据集进行调节）。计算测试属性组的属性值的平均值$\frac{1}{k}\sum_{i=1}^{n=k} a_i$，并将其作为最佳分裂点进行划分。

连续属性的离散化避免了连续属性的多次分裂，从而减少了森林的结点数目和空间开销。

改进策略 1 的具体步骤描述如下：

步骤 1　根据数据集输入的属性描述，确定数据集中的连续属性。

步骤 2　对连续属性进行排序，并确定某一连续型属性的相邻两类边界区。

步骤 3　随机抽取边界区处的 k 个连续属性值$\{a_1, a_2, \cdots, a_k\}$作为测试属性组，计算测试属性组的属性值的平均值$\frac{1}{k}\sum_{i=1}^{n=k} a_i$。

步骤 4　将测试属性组的属性值平均值作为最佳分割阈值进行划分。

2.2 改进策略 2：弱相关化特征子空间

通过对随机森林算法的分析可知，决策树之间的相关性越小，随机森林分类效果越好[12]。因此本文通过采用相关性检测的方法，在保证随机抽样的前提下，选取与已建树属性相关性最小的特征子空间作为参与建树的属性，从而提高决策树对领域的覆盖精度，降低决策树间的相关性，进而改善分类效果。

2.2.1 属性向量空间模型

本文结合统计学方法中的向量空间模型，提出了属性向量空间模型。在采用该模型进行属性之间的相关性计算时，需要计算属性值在属性中的重要程度，计算时一般采用 TF-IDF 方法，在使用该方法计算属性值的权重时会涉及到两个概念。

（1）属性频率 $TF_{P,T}$

属性频率 $TF_{P,T}$ 是指一个属性值在数据集某个属性列中出现的频率。对于一个属性值来讲，我们用属性频率这个概念来表示该属性值在某属性列中所占的比重，计算公式为：

$$TF_{P,T}=\frac{i_{P,T}}{\sum_{L} i_{L,T}} \tag{1}$$

其中，$i_{P,T}$ 为该属性值在某属性 T 列中的出现次数，而 $\sum_{L} i_{L,T}$ 则为某属性 T 列中所有属性值的出现次数之和。

例如“Private”这一属性值在该属性列中出现的次数为 5，而该属性列中属性值数为 100，则 $TF_{Private,T}=\frac{5}{100}=0.05$。

（2）逆向属性频率 IDF_P 是指一个属性值的逆向属性频率，这也是一个比较重要的权重计算方法。其计算公式为：

$$IDF_P=\log\frac{|D|}{1+|\{T:P_m\in D_n\}|} \tag{2}$$

其中，$|D|$ 为数据集中属性的总数，$|\{T:P_m\in D_n\}|$ 为包含属性值 P_m 的属性个数。

例如“Private”这一属性值在数据集 5 个属性中出现过，而数据集的属性总数是 20，则 $IDF_{Private}=\frac{25}{5}=5$。

因此这种加权方法可以表示为 $TF*IDF(Private,T)=TF_{Private,T}*IDF_T$。上述的例子可以表述成属性 T 中属性值“Private”的属性频率 * 逆向属性频率即 $0.05\times5=0.25$。

2.2.2 属性之间的相关性计算

由于一列属性可以由属性值来表示，因此属性之间的相关性可以由属性值向量之间的相关性来描述。

设 v_i, v_j 是两列不同属性的属性值向量：

$v_i=(TF*IDF(a_1,i), TF*IDF(a_2,i), \cdots, TF*IDF(a_k,i))$

$v_j=(TF*IDF(a_1,j), TF*IDF(a_2,j), \cdots, TF*IDF(a_k,j))$

因此，两列属性 i,j 之间的相关性如式(3)：

$$r(v_i,v_j)=\frac{\sum_{n=1}^{k} TF*IDF(a_1,i)\times TF*IDF(a_1,j)}{\sqrt{\sum_{n=1}^{k}(TF*IDF(a_1,i))^2}\times\sqrt{\sum_{n=1}^{k}(TF*IDF(a_1,j))^2}} \tag{3}$$

式(3)表示向量的 v_i 和 v_j 之间的相关度，该计算值越大表示两属性之间的相关性越强。

2.2.3 选择弱相关化特征子空间

通过式(3)计算抽样属性与已建树属性的相关性的均值 $\bar{r}$，$\bar{r}$ 越接近 0 表示相关性越小。在 $2n$ 个随机抽样属性中，选择与已建树属性相关性最小的 m 个属性作为特征子空间，从而达到特征子空间弱相关化的目的。

改进策略 2 的具体步骤描述如下：

步骤 1　为当前正在构建的决策树随机抽取 $2n$ 个属性。

步骤 2　构建中的决策树读取已构建的决策树对应结点的属性信息。

步骤 3　计算抽样属性与已建树结点属性之间的相关性。

步骤 4　选择相关性均值最小的 m 个抽样属性，构造弱相关化特征子空间。

3 算法设计方案

3.1 DWRF 算法森林构建阶段并行化策略

随机森林是由多棵决策树$\{h(X,\theta_k),k=1,2,\cdots,K\}$组成的分类器，其中$\{\theta_k\}$是相互独立且同分布的随机向量。$K$ 表示随机森林中决策树的棵数，最后由全部决策树分类器通过投票确定输入向量 X 的最终分类标签[2]。由此可见，每棵决策树的建立均不依赖于其他决策树，决策树与决策树之间互相独立。因此，随机森林的算法原理为 DWRF 算法的并行化提供了理论依据，且其结构十分合适在 Hadoop 集群上进行线性扩展。

通过对随机森林算法的进一步分析，在建立决策树的过程中需要计算每个结点的分裂属性，此过程需要将特征子空间中所有的属性进行遍历，计算每个属性相应的信息增益率。考虑到信息增益率计算是基于属性间相互独立的特点，此过程可以进一步并行化。可采用多线程的方法，使每个子线程完成遍历特征子空间及计算信息增益率的过程后，将子线程的结果汇总，得到最优分裂属性。结合上节的改进策略 1 和改进策略 2，本文 DWRF 算法的并行化流程如图 1 所示，具体步骤如下。

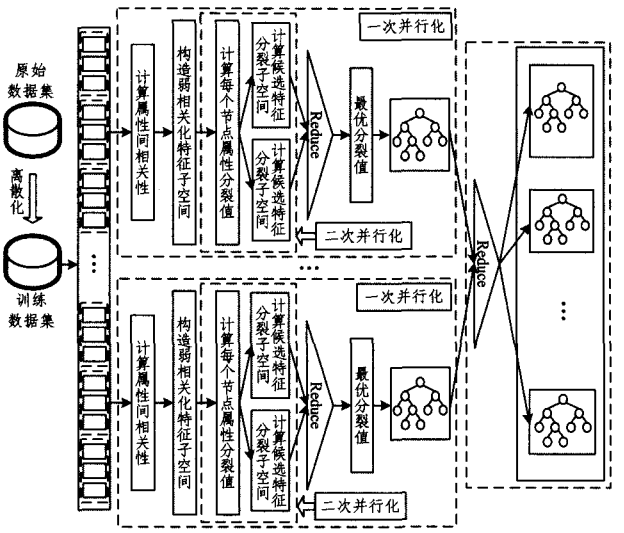

图 1 基于 MapReduce 的 DWRF 算法的并行化流程

输入：训练数据集 DS，决策树棵数 n，弱相关化特征子空间所包含属性个数 m

输出：DWRF 森林模型

步骤 1 对训练数据集 DS 进行预处理，若 DS 中包含连续属性，利用改进策略 1 对其进行离散化；

步骤 2 利用 bootstrap 方法抽取每棵决策树的样本子空间，构成该决策树的训练样本集；

步骤 3 将每棵决策树的训练样本集分配到不同的计算节点上；

步骤 4 根据上文的改进策略 2 为每棵决策树选择弱相关化特征子空间；

步骤 5 划分弱相关化特征子空间，作为每个子线程处理的候选分裂特征子空间；

步骤 6 每个子线程计算各自对应的候选分裂特征子空间中所有属性值的信息增益率，返回最优的信息增益率和对应的属性名称，如〈〈TrID，nodeID，FeatureID〉，value〉；

步骤 7 Reduce 子线程返回的结果，选取分裂值最优的作为该结点的分裂属性和分裂点；

步骤 8 重复步骤 4－步骤 8，直到满足终止迭代的条件，完成决策树的构建；

步骤 9 将决策树输出到 HDFS 上，进行组合集成，完成 DWRF 森林模型的构建。

3.2 DWRF 算法投票阶段并行化策略

在传统的随机森林算法中，每棵决策树依次投票给测试数据集样本，例如，第一棵决策树完成对样本的投票以后，第二棵决策树才能开始对这个样本进行投票。每棵决策树投票过程相互独立。故此阶段也十分适于并行化。

基于 MapReduce 的 DWRF 算法投票阶段并行化实现流程如图 2 所示，具体操作步骤如下。

输入：测试数据集 DT

输出：分类结果

步骤 1 将每棵决策树 Map 到相应的计算节点；

步骤 2 每棵决策树分配一个 Map 函数，当输入测试样本时，预测该样本，返回投票结果〈Tm，〈$Tr\ ID$，$Label$〉〉；

步骤 3 调用 Reduce 函数，统计步骤 2 中返回的投票结果；

步骤 4 选择票数最多的类别作为该测试样本的分类预测类别。

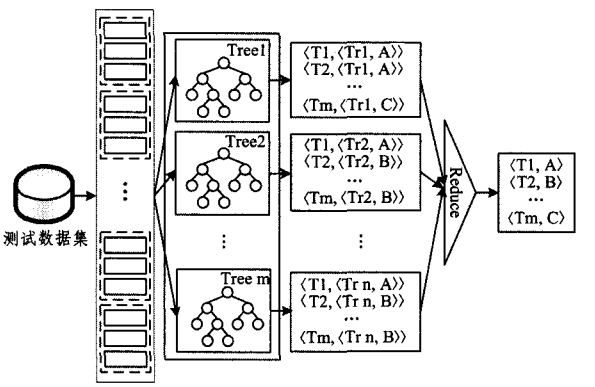

图 2 基于 MapReduce 的 DWRF 算法投票阶段的并行化流程

4 实验与结果分析

4.1 实验环境

图 3 为实验中 Hadoop 分布式计算集群的结构，1 台 PC 机作为 NameNode 与 Job Tracker 服务节点，4 台其他的 PC 机作为 DataNode 与 Task Tracker 服务节点。每台节点硬件配置如下：CPU 型号为 Inter(R)Core(TM)i7-4790M，内存为 8GB，硬盘为 1TB，Hadoop 版本为 1.2.1。

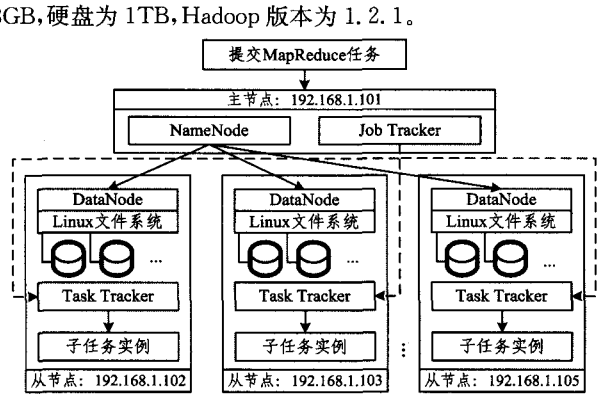

图 3 Hadoop 分布式计算集群结构

4.2 实验数据集

本实验选取 UCI 标准数据集 Glass Identification Data

Set(简称 Glass)，MAGIC gamma telescope data 2004(简称 MAGIC)，Poker Hand，Covertype Data Set(简称 Covertype)来测试算法的基本性能，标准数据集描述详见表 1。其余实验选取某绿色数据中心的日志数据作为实验数据集(数据集 FX)，该数据集描述详见表 2。

表 1　标准数据集相关描述

序号	数据集名称	实例数	属性数	类别数	数据集规模
1	Glass	214	10	7	12k
2	MAGIC	19020	11	2	1.4M
3	Poker Hand	1025010	11	9	23M
4	Covertype	581012	54	7	71.6M

表 2　实验数据集 FX 相关描述

序号	数据集名称	实例数	属性数	类别数	数据集规模
1	FX_1	5253552	23	8	0.8G
2	FX_2	7155030	23	8	1.2G
3	FX_3	10507104	23	8	1.6G
4	FX_4	14310060	23	8	2G

4.3　算法基本性能测试实验

本实验首先分别从 Glass，MAGIC，Poker Hand，Covertype 数据集抽取 70%作为训练集，余下的 30%作为测试集。然后分别采用以下 3 种方法建立分类模型：方法 1，采用随机森林算法(RF)直接建立分类模型；方法 2，采用离散化随机森林算法(DRF)，首先对数据集中的连续属性进行离散化处理，然后建立分类模型；方法 3，采用 DWRF 算法，在方法 2 的基础上，将弱相关化特征子空间运用到建立决策树的属性选择阶段，并建立相应的分类模型。对方法 1－方法 3 得到的每个分类模型都使用对应的数据集进行多次测试，取结果的平均值作为最终实验结果。所得的分类准确率如图 4－图 7 所示。

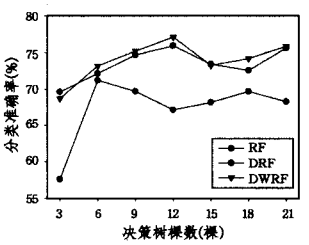

图 4　Glass 数据集分类准确率

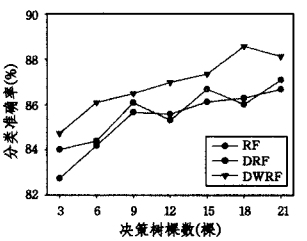

图 5　MAGIC 数据集分类准确率

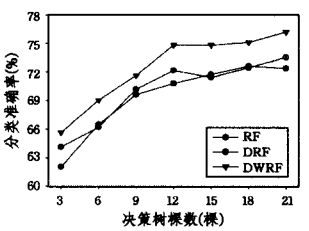

图 6　Poker Hand 数据集分类准确率

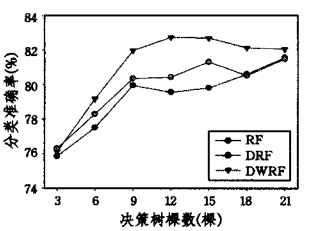

图 7　Covertype 数据集分类准确率

实验结果显示，3 种算法的准确率都随着森林中决策树棵数的增加而提高，当森林中的决策树达到一定数量时，准确率开始趋于稳定。离散化随机森林算法(DRF)和随机森林算法(RF)在除数据集 Glass 之外的其它数据率上准确率相当。在 Glass 数据集上 DRF 算法的准确率要明显高于 RF 的，这是因为 Glass 数据集目标属性分布比较集中，利于发挥改进策略 1 的优势。而 DWRF 算法比 DRF 算法和 RF 算法的准确率都要高，并且在数据对象较多的 Poker Hand 数据集和属性维数较多的 Covertype 数据集上也有较好的表现。

表 3 列出了不同分类算法在最优参数(决策树数目)下生的成结点数量，DRF 和 RF 的结果显示，数据离散化处理(改进策略 1)能在保证分类准确率相当的情况下减少树的结点个数，生成规模更小的森林。而通过分析 DWRF 的实验结果可知弱相关化特征子空间的选取(改进策略 2)不会促使森林规模的增大。

表 3　不同分类算法在最优参数下生成结点对比

测试算法	实验数据集			
	Glass	MAGIC	Poker Hand	Covertype
	森林平均结点个数			
RF	39	1541	256818	9470
DRF	42	1507	254039	9095
DWRF	38	1492	254123	8923

为了证明该算法在大规模数据集上也有较好的性能，本文利用数据集 FX1，FX2，FX3，FX4 对其分类准确性进行了测试。实验结果如图 8 所示。

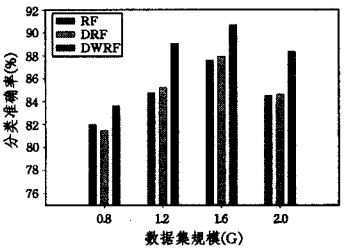

图 8　分类准确率随数据集规模改变的变化情况

图 8 中的实验结果显示，对于不同规模的大规模数据集(0.8G，1.2G，1.6G，2.0G)，本文提出的 DWRF 算法建立的分类模型比 RF 和 DRF 算法生成的分类模型在分类准确率上更高。因此 DWRF 算法也适用于大规模数据的分类。

4.4　集群的加速比实验

此部分实验主要考虑两个方面的问题[13]：(1)当处理相同规模的数据且集群规模不断增大时，Hadoop 平台的分布式 DWRF 并行算法构建森林的能力。(2)当数据和集群规模都呈相同比例增长和减少时，Hadoop 平台的分布式 DWRF 并行算法构建森林的能力。

对于第一点，分别在 1、2、3、4、5 台集群上进行实验，实验数据为实验数据集 FX1(0.8G)、FX2(1.2GB)、FX3(1.6GB)、FX4(2.0G)，在不断增加集群数目的过程中，分别记录每次算法的运行时间，得到如图 9 所示的结果。从图中可以看出，算法运行的时间随着集群规模的不断增大而减少，当集群规模相同时，数据集越小，运行的速度越快。所以，增大集群的规模可以明显提高 DWRF 处理相同大小数据集的效率，这说明 DWRF 并行算法适合处理大规模数据集。

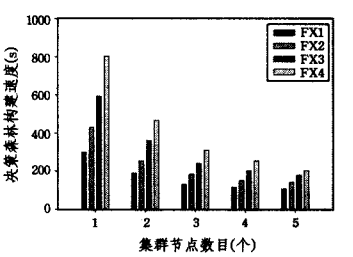

图 9　同样规模的数据，集群增大时的随机森林构建速度

(下转第 90 页)

bedded VoIP terminal based on LINPHONE[J]. Information Communication,2013(8):77-79(in Chinese)
沙爱军,沈卫康,毛其林. 基于 LINPHONE 的嵌入式 VoIP 终端实现[J]. 信息通信,2013(8):77-79

[5] Zhao Xin. Design and implementation of call control model based on SIP protocol[D]. Hefei: Hefei University of Technology, 2010(in Chinese)
赵昕. 基于 SIP 协议的呼叫控制模型设计与实现[D]. 合肥:合肥工业大学,2010

[6] Lu Hua, Wang Bao-bao. Research and application of oSIP protocol stack[J]. Electronic Science and Technology, 2006(2):61-64(in Chinese)
卢华,王保保. oSIP 协议栈的研究及应用[J]. 电子科技,2006(2):61-64

[7] Liu Xi-yi. Research and implementation of VoIP soft terminal based on SIP protocol[J]. Information Security and Technology, 2011(10):33-35(in Chinese)
刘习义. 基于 SIP 协议 VoIP 软终端的研究与实现[J]. 信息安全与技术,2011(10):33-35

[8] Li Zhen-jun, Zeng Ling-yun. Research and implementation of embedded SIP terminal[J]. Manufacturing Automation, 2011, 33(7):141-144(in Chinese)
李振军,曾凌云. 嵌入式 SIP 终端的研究与实现[J]. 制造业自动化,2011,33(7):141-144

(上接第 58 页)

对于第二点,本文分别使用 2、3、4、5 台节点的集群来处理数据集 FX1、FX2、FX3、FX4。从图 10 中可以看出,当集群规模和数据集大小都呈比例增长时,Hadoop 平台对数据的处理能力基本持平,这两点都充分体现了该算法在 Hadoop 平台上具有较强的扩展性。

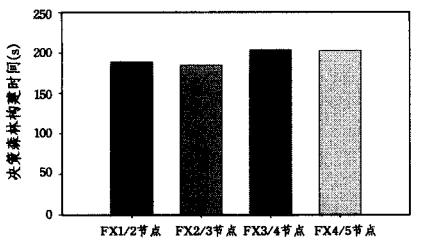

图 10 数据和资源同比例增长和减少时随机森林构建速度

结束语 传统的分类算法在面对大规模数据集时,显得力不从心。本文在已有研究成果的基础上,提出了一种基于弱相关化特征子空间选择的离散化随机森林并行分类算法。实验证明,DWRF 算法不仅能有效提高分类准确率,降低森林的空间开销,而且在分布式环境下有着良好的并行性和扩展性,因此说明本文所采用的方法是有效的。

参 考 文 献

[1] HE Qing, LI Ning, LUO Wen-Juan, et al. A Survey of Machine Learning Algorithms for Big Data [J]. Pattern Recognition and Artificial Intelligence, 2014, 27(4):327-336(in Chinese)
何清,李宁,罗文娟,等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能,2014,27(4):327-336

[2] Breiman L. Random forests [J]. Machine learning, 2001, 45(1): 5-32

[3] Wang Y, Goh W, Wong L, et al. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes [J]. BMC bioinformatics, 2013, 14(16):1-15

[4] Zhang Lei, Wang Lin-lin, Zhang Xu-dong, et al. The basic principle of random forest and its applications in ecology: a case study of Pinus yunnanensis [J]. Acta Ecologica Sinica, 2014, 34(3): 650-659(in Chinese)
张雷,王琳琳,张旭东,等. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例[J]. 生态学报,2014,34(3):650-659

[5] Lee S L A, Kouzani A Z, Hu E J. Random forest based lung nodule classification aided by clustering [J]. Computerized Medical Imaging and Graphics, 2010, 34(7):535-542

[6] Luo Zhi-lin, Chen Ting, Cai Wan-dong. Microblogging Retweet Prediction Algorithm Based on Random Forest[J]. Computer Science, 2014, 41(4):62-64, 74(in Chinese)
罗知林,陈挺,蔡皖东. 一个基于随机森林的微博转发预测算法[J]. 计算机科学,2014,41(4):62-64,74

[7] Wang De-wen, Sun Zhi-wei. Big Data Analysis and Parallel Load Forecasting of Electric Power User Side[J]. Proceedings of the CSEE, 2015, 35(3):527-537(in Chinese)
王德文,孙志伟. 电力用户侧大数据分析与并行负荷预测[J]. 中国电机工程学报,2015,35(3):527-537

[8] Guo Shan-qing, Gao Cong, Yao Jian, et al. An Intrusion Detection Model Based on Improved Random Forests Algorithm [J]. Journal of Software, 2005, 16(8):1490-1498(in Chinese)
郭山清,高丛,姚建,等. 基于改进的随机森林算法的入侵检测模型[J]. 软件学报,2005,16(8):1490-1498

[9] Yao Dong, Luo Jun-yong, Chen Wu-ping, et al. Online Double Random Forests Intrusion Detection Based on Non-extensive Entropy Features Extraction [J]. Computer Science, 2013, 40(12):192-196(in Chinese)
姚东,罗军勇,陈武平,等. 基于改进非广延熵特征提取的双随机森林实时入侵检测方法[J]. 计算机科学,2013,40(12):192-196

[10] Hu Qing, Sun Cai-xin, Du Lin, et al. Transformer Fault Diagnosis Method Using Random Forests and Kernel Principle Component Analysis [J]. High Voltage Engineering, 2010, 36 (7): 1725-1729(in Chinese)
胡青,孙才新,杜林,等. 核主成分分析与随机森林相结合的变压器故障诊断方法[J]. 高电压技术,2010,36 (7):1725-1729

[11] Yao Ya-fu, Xing Liu-tao. Improvement of C4. 5 decision tree continuous attributes segmentation threshold algorithm and its application [J]. Journal of Central South University (Science and Technology), 2011, 42(12):3772-3776(in Chinese)
姚亚夫,邢留涛. 决策树 C4.5 连续属性分割阈值算法改进及其应用[J]. 中南大学学报(自然科学版),2011,42(12):3772-3776

[12] Xu B, Huang J Z, Williams G, et al. Classifying very high-dimensional data with random forests built from small subspaces[J]. International Journal of Data Warehousing and Mining (IJDWM), 2012, 8(2):44-63

[13] Xiang Yao, Yuan Jing-ling, Zhong Luo, et al. A Coarse-Grained Clustering Unit Based Parallel Algorithm for Big Data Set [J]. Journal of Chinese Computer Systems, 2014, 35(10):2370-2374(in Chinese)
向尧,袁景凌,钟珞,等. 一种面向大数据集的粗粒度并行聚类算法研究[J]. 小型微型计算机系统,2014,35(10):2370-2374