文章编号: 1004-4353(2017)01-0055-05

基于 MapReduce 的朴素贝叶斯算法在新闻分类中的应用

徐保鑫, 怀丽波*, 崔荣一

(延边大学工学院 计算机科学与技术学科 智能信息处理研究室, 吉林 延吉 133002)

摘要: 针对传统单点串行的分类算法在面对新闻数据规模较大、分类属性较多时存在效率低的问题,本文研究了朴素贝叶斯分类算法在 MapReduce 下的并行实现方法. 首先对新闻信息进行分词、格式转换等预处理,然后进行特征提取、分类模型构造;最后进行了分类测试. 测试结果表明,在大数据量的情况下,并行化的贝叶斯算法较传统的贝叶斯算法具有更好的执行效率和较高的扩展性.

关键词: Hadoop; 朴素贝叶斯; MapReduce; 文本分类; 新闻文本

中图分类号: TP391.3 **文献标识码**: A

Naive Bayes algorithm application in the classification of news based on MapReduce

XU Baoxin, HUAI Libo*, CUI Rongyi

(*Intelligent Information Processing Lab., Dept. of Computer Science & Technology, Yanbian University, Yanji* 133002, *China*)

Abstract: According to the traditional single point serial classification algorithm in the face of the existence of the problem of low efficiency, large scale news data classification attribute more, in this paper naive Bayesian classification algorithm in MapReduce parallel implementation method. First of all, the word segmentation and format conversion are processed, then the feature extraction and classification model are constructed. The test results show that, in the case of large amount of data, the parallel Bayesian algorithm has better performance and scalability than the traditional Bayesian algorithm.

Keywords: Hadoop; naive Bayes; MapReduce; text classification; news text

0 引言

新闻网站的分类导航是将新闻资源按照一定的体系组合,给用户提供各级类目,方便用户浏览检索[1],但面对爆炸式的信息增长速度,用户获取准确信息的难度越来越加大,因此迫切需要对新闻信息进行有效的整理. 文本分类技术是信息组织、文本挖掘的重要基础,可以较大程度地解决信息紊乱的问题,帮助用户准确地定位所需的信息,是目前处理海量信息的重要手段. 针对海量新闻信息整理问题,有学者将文本分类技术引人到该领域,例如:李安将国外的 Factiva 分类标引体系引入到我国的新闻分类中[2];国内新华社等媒体采用归类技术进行新闻分类[3];张志平所采用的是中文新闻信息分类体系[4],张永奎等人的研究中采用了三层突发事件新闻分类体系[5]. 但这些都是针对小规模的数据量,在处理海量新闻时间仍需要花费过长的时间,因此文本采用 Hadoop

收稿日期: 2016-12-15

* **通信作者**: 怀丽波(1973—),女,副教授,研究方向为优化理论与方法、数据挖掘.

框架,对新闻进行并行条件下的分类计算,从而从时间上对新闻进行了优化.

1 朴素贝叶斯概述及 MapReduce 框架简介

1.1 朴素贝叶斯算法概述

朴素贝叶斯分类方法是目前最常用的分类算法之一,具有所需估计的参数很少,对缺失数据不太敏感的特点.应用此分类方法得到的分类效果很理想,并且计算效率也比较高[6].因此在本文中采用此算法构造分类器并对新闻进行分类.该分类器的基本原理是先通过总结经验获得先验概率,再根据贝叶斯公式计算出所对应的后验概率,得出这个对象属于某个类别的概率,然后选择后验概率最大的类别作为该对象的最终所属类别[7].朴素贝叶斯算法的基本描述如下:

设有数据集 T,共分成 N 类 $C_1,C_2,\cdots,C_N$,数据集中的每个样本有 n 个属性,$X_i=\{1,2,\cdots,n\}$ 表示第 i 个属性,对于一个给定的待分类样本 X 和类别 $C_i(1\leqslant i\leqslant N)$,朴素贝叶斯分类算法使用最大的后验概率 $P(C_i|X)$ 来预测 X 所属的类别.根据贝叶斯定理

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}. \tag{1}$$

先验概率 $P(C_i)$ 是指根据以前经验分析得到的概率,比如说全概率公式,它一般是作为"由因求果"问题中的"因"出现,这类概率是检验前概率,并没有进行试验验证,所以称为先验概率.

$$P(C_i)=n_i/N, \tag{2}$$

其中:n_i 为类别 C_i 的训练样本数,N 为训练样本的总数.

条件概率是指如果在事件 B 已经发生的条件下考虑事件 A 发生的概率,就称它为事件 A 的条件概率,表示为 $P(A|B)$.按照原理将朴素贝叶斯分类应用到文本分类中得到:

$P(X|C_i)=$(类别 C_i 中样本 X 的属性在各个文档中出现过的次数之和 $+1$)/(类别 C_i 中的单词总数 $+$ 训练样本中不相同的特征词总数). (3)

由于朴素贝叶斯假设各个属性之间是相互独立的,因此有

$$P(X|C_i)=P(X_1|C_i)P(X_2|C_i)\cdots P(X_n|C_i)=\prod_{j=1}^{n}P(X_j|C_i). \tag{4}$$

对于任意未知分类的样本 X,当且仅当下面条件是

$$P(C_i|X)P(C_i)>P(C_p|X)P(C_p)\quad(1\leqslant p\leqslant N,\ p\neq i), \tag{5}$$

将 X 归为 C_i 类.

凡事都有两面性,NBC 算法也存在一定的不足.在实际应用中基本上不可能满足其属性条件独立性的假设,对那些属性间存在高度相关性的数据,如果直接使用 NBC 进行处理,分类效果很难达到实际预期[8].另外,在需要处理的数据不完整,或者出现极度不平衡数据时可能就会导致某个甚至某些属性的后验概率出现较大偏差,从而影响最终的分类结果.不过,目前已有相关方法用以解决数据不完整和不平衡数据问题,比如拉普拉斯平滑技术、属性加权方法等.这些方法在一定程度上可以提高 NBC 的性能.

1.2 MapReduce 编程框架简介

MapReduce 由普通 PC 机群构成,采用"分而治之"的思想,把对大规模数据集的操作,分发给一个主节点管理下的各个分节点共同完成,然后通过整合各个节点的中间结果,得到最终结果[9].其主要思想是将要执行的问题自动拆解成 map(映射)和 reduce(化简)的方式.MapReduce 框架运转在〈key,value〉键值对上,也就是说,框架把作业的输入看作一组〈key,value〉,同样也产生一组〈key,value〉作为作业的输出,这两组键值对的类型可能不同.处理时,每个节点就近读取本地存储的数据处理(Map),将处理后的数据进行合并(combine)再分发至 Reduce 节点,避免了大量数据的传输,提高了处理效率[10].一个 MapReduce 作业的输入和输出类型如下所示:(input)〈key1,value1〉→ map→ 〈k2,v2〉→ combine→ 〈k2,v2〉→ reduce→ (output)〈k3,v3〉.

MapReduce 分布式处理框架,不仅可处理大规模数据,而且能将很多繁琐的细节隐藏起来,比如自动并行化、负载均衡等,并且 MapReduce 的伸缩性非常好,每增加一台服务器,就能将该服务器的计算能力接入到集群中,可以极大地节约成本[11].

2　基于 MapReduce 的朴素贝叶斯模型构造

朴素贝叶斯算法是利用统计学知识进行分类的算法，在样本各属性之间条件独立的前提下，该算法表现出较高的分类准确率[12]. 但因为其传统串行的实现方式时间复杂度较高，处理海量数据具有局限性[13]. 针对传统算法的无法处理或者无法高效处理海量数据的缺陷[14]，本文通过研究 MapReduce 基本框架和朴素贝叶斯算法，在 Apache提供 Hadoop 分布式文件系统（HDFS）和 MapReduce 并行计算框架的基础上，研究了朴素贝叶斯的并行化算法实现. 通过将大的新闻数据集分解成多个子数据集分配给多个节点并行处理的方式，减少构造分类模型和测试的时间，提高对新闻数据分类的效率.

本文将贝叶斯分类算法在 MapReduce 框架下并行化的过程分为 4 个阶段：数据预处理阶段、特征提取阶段、模型训练阶段和测试阶段，具体流程如图 1 所示.

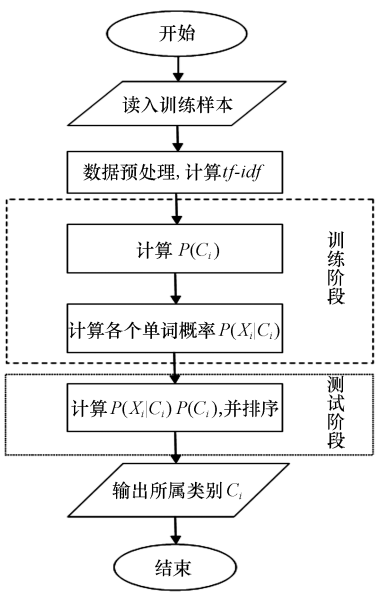

图 1　朴素贝叶斯算法流程图

2.1　数据预处理阶段

输入：训练样本集.

输出：序列文件 chunk-0.

1）原始的新闻数据是一堆分类目录，每个目录名就是类别名，目录里包含属于此类的文档，将这些文档合并到一个序列文件 chunk 中，存储类型是（key：Text，value：Text），其中每个键值对的 key 代表其对应类别（目录）名，value 代表该目录下的新闻信息的内容.

2）对 chunk 文件中的数据进行分词处理，去掉标点及副词，并通过停用词的方式去除虚词和无意义词等，方便计算机进行识别与计算，保存分词后的文件 chunk-0.

2.2　特征提取阶段

特征提取模块是以分词后的新闻列表为输入，根据特征提取算法选出特征向量. 本实验采用的是 TF-IDF 算法，具体步骤为：

Step1　统计相关数据.

输入：chunk-0 文件.

输出：wordFreq（词频 TF 值）；termDocCount（类 C_i 出现特征词的文档数）；featureCount（训练集出现特征词的总数）；docCount（每个类的文档数）.

Map：对 chunk-0 文件进行分块，计算各个分块的 TF、类 C_i 出现特征词的文档数、训练集出现特征词的总数以及每个类型的文档数.

Reduce：将 Map 的输出求和，将小于最小词频数的词过滤，并把结果输出.

Step2　计算 $tf\text{-}idf$.

输入：Step1 中的输出文件.

输出：vocabcount 文件（训练样本中不相同的特征词总数），tf-idf 文件.

Map：1. 提取 TF，直接输出.

2. 计算 IDF＝Log（类别下的文档数/词的文档数）；

3. 计算不相同的特征词数量.

Reduce：1. 对于特征词数量，求和；

2. 计算 $tf\text{-}idf = TF * IDF$.

2.3　分类模型训练阶段

在新闻样本训练阶段，根据朴素贝叶斯分类算法思想，对经过预处理与特征提取后的特征词集合计算每个类别在训练样本中的出现频率 $P(C_i)$ 及每个特征属性划分对每个类别的条件概率 $P(X_i|C_i)$ 构成分类器的参数. 具体流程为：

Step1　利用公式（2）计算每个类别的先验概率 $P(C_i)$，$P(C_i)$＝类 C_i 的训练样本数 / 训练样本总数.

Step2　计算类别和特征的权重.

输入:tf-idf 文件.

输出:weights 文件.

Map:1.从 key 中拿到 feature 输出其 *tf-idf* 值,计算每个属性的 *tf-idf* 特征权重和;

2.从 key 中拿到 lable 输出其 *tf-idf* 值,计算某类下的全部属性的 *tf-idf* 权重和;

3.计算全部的特征所有类型的权重和.

Reduce:对 map 的结果求和,输出权重文件 weights.

Step3 利用公式(3)计算先验概率,即 $P(X_i|C_i)$.

输入:tf-idf 文件,vocabcount 文件.

输出:Normalizer 文档.

Map:对输入的每一个特征词计算 Weight = (tf-idf + 1.0)/(sigma_k + VocabCount)
//sigma_k 为某类下的所有属性的 tf-idf 总和,由 Step2 得到.

Reduce:对 Map 结果求和.

2.4 分类模型测试阶段

根据公式(1)可以得到测试样本分别属于每一类的概率,即 $P(C_i|X)$,对输出的结果进行从大到小的排序,可得到该样本数据的所属类别.

输入:测试训练集.

输出:分类结果.

Map:计算测试样本的 *tf-idf*,权重 *weight*.计算其后验概率,输出当前最大后验概率值,即确定测试样本类别.

Reduce:合并 Map,输出最终分类结果.

3 实验结果及分析

为了验证并行朴素贝叶斯算法的准确度及效率,本文搭建了一个 Hadoop 集群,由 1 个主节点 master 和 3 个从节点 slave 构成.计算机配置为 IntelCorei5-3470CPU@3.20 GHz,内存为 8 GB,系统为 CentOS6.3.实验数据集来自 Hadoop 应用开发实战案例中的网络新闻信息,分为 MP3、camera、computer、household、mobile 5 种类别,文本规模为 1 万篇文档,实验过程将新闻数据按 6∶4 分为训练集和测试集.

第 1 组实验是加速比性能实验.分别选择 1、2、3 个节点对上述数据集进行交叉分类测试实验,记录包括文本预处理、分类模型计算、测试在内的总处理时间,如表 1 所示.从表 1 可以看出,随着节点数的增加,处理相同数据的运行时间减少,由此说明数据集较多时,可以通过增加节点数的办法来提高算法的执行速度.

表 1 不同节点下的时间对比图

节点数	运行时间/s
1	105
2	93
3	64

第 2 组实验是对 Hadoop 下的分类模型性能进行测试,使用准确率来进行评估.数据测试结果一共分为 4 种情况:*TP*(预测为正,实现为正);*FP*(预测为正,实现为负);*FN*(预测为负,实现为正);*TN*(预测为负,实现为负).准确率的计算公式如下:

$$\text{准确率(precision)} = \frac{TP}{TP + FP}. \tag{6}$$

输出的矩阵为分类器对测试样本的输出结果,结果如表 2 所示.根据公式(6)计算得出分类的总准确率为 96.27%.从分类精度上看,手机类的分类精度最低,其大多误判为电脑类.

表 2 测试输出的矩阵

类别	MP3	camera	computer	household	mobile	总数
MP3	618	2	2	5	16	643
camera	3	493	0	17	2	515
computer	12	2	540	10	37	601
household	0	3	0	589	4	596
mobile	4	3	0	1	940	948

通过对矩阵进行分析可知，影响分类准确率的因素可能是如下几个方面：1）受所选的中文分词工具性能的影响，在本研究中没有进行专门的特征词选取工作，所以分词工具所提供的停用词功能对实验结果有较大的影响. 2）与语料库中样本质量有关，每一个类之间应该存在较大的区分度.

第 3 组实验是将传统的朴素贝叶斯算法和并行化的贝叶斯算法进行比较，其结果如图 2 所示. 从图 2 可以看出：在处理少量的新闻数据集时，串行贝叶斯算法的效率优于并行的贝叶斯算法，这是因为 MapReduce 的 job 调度以及对数据集进行分块和重组等过程都需要一定的时间；但随着新闻规模的不断增长，并行贝叶斯分类算法的效率逐步优于串行贝叶斯算法，且数据规模越大优势越明显.

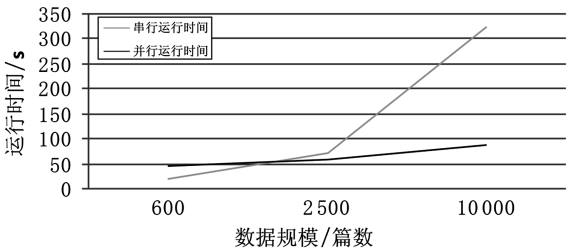

图 2　传统的朴素贝叶斯算法和并行化的贝叶斯算法的运行时间

4　结束语

本文对朴素贝叶斯算法及 MapReduce 框架进行了简要的分析，并利用 Hadoop 平台搭建了基于 MapReduce 编程框架下的朴素贝叶斯的并行算法，通过实验验证了当数据较大时并行的分类算法效率比传统的串行算法高. 实验表明，对大规模的新闻使用并行化的分类算法是一种有效的分析处理方法. 但本文由于实验材料以及实验环境所限，本文仅利用实验室有限的几台 PC 搭建了实验环境，语料库的规模也比较小，所以在今后的研究中，我们将适当增加集群的节点数和扩大数据规模，以更好地测试本文算法的分类效果.

参考文献：

[1] 喻国明，李彪. 新闻传播的大数据时代[M]. 北京：中国人民大学出版社，2014.

[2] 李安. Factiva 新闻分类标引体系及其对我国的启示[J]. 图书馆建设，2003(3)：102-104.

[3] 百度百科. 新华网[EB/OL]. [2013-04-18]. http://baike.baidu.com/view/154954.htm.

[4] 张志平. 基于“中文新闻信息分类与代码”文本分类[J]. 太原理工大学学报，2010(4)：402-405.

[5] 张永奎，李红娟. 基于类别关键词的突发事件新闻文本分类方法[J]. 计算机应用，2005(51)：139-140.

[6] 马宾，殷立峰. 一种基于 Hadoop 平台的并行朴素贝叶斯网络舆情快速分类算法[J]. 现代图书情报技术，2015(2)：78-84.

[7] 段晶. 朴素贝叶斯分类及其应用研究[D]. 大连：大连海事大学，2011.

[8] Jiang Liangxiao，Li Chaoqun，Wang Shasha，et al. Deep feature weighting for naive bayes and its application to text classification[J]. Engineering Applications of Artificial Intelligence，2016，52：26-39.

[9] Tom White. Hadoop 权威指南[M]. 2 版. 北京：清华大学出版社，2011：15-73，167-188.

[10] 李伟卫，赵航，张阳. 基于 MapReduce 的海量数据挖掘技术研究[J]. 计算机工程与应用，2013，49(20)：112-117.

[11] 朱珠. 基于 Hadoop 的海量数据处理模型研究与应用[D]. 北京：北京邮电大学，2008.

[12] 李方，刘琼荪. 基于改进属性加权的朴素贝叶斯分类模型[J]. 计算机工程与应用，2010(4)：132-133.

[13] 郭绪坤，范冰冰. 一种朴素贝叶斯文本分类算法的分布并行实现[J]. 计算机应用与软件，2016(11)：240-243.

[14] 严嘉铭，黄理灿. 基于 MapReduce 的朴素贝叶斯文本分类研究[J]. 工业控制计算机，2016，29(4)：96-97.