



基于云平台的互信息最大化特征提取方法研究^{*}

魏莎莎¹, 陆慧娟¹, 金伟¹, 李超²

(1. 中国计量学院信息工程学院 杭州 310018; 2. 中国计量学院机电工程学院 杭州 310018)

摘要:随着大规模基因芯片的应用, 针对高维度的基因表达数据存在大量无关和冗余特征可能降低分类器性能的问题, 提出了一种基于云平台的互信息最大化特征提取(CMI-Selection)方法。Hadoop 云计算平台对基因表达数据划分后进行并行计算, 同时结合互信息最大化方法对特征进行提取, 实现了云计算平台上的特征过滤模型。实验结果表明, 基于云平台的互信息最大化特征提取方法能够在保证较高分类精度的情况下, 快速提取特征, 节省大量时间资源, 是一种高效的基因特征提取系统。

关键词: 云计算; 互信息最大化; 云平台; 特征提取

doi: 10.3969/j.issn.1000-0801.2013.10.008

Maximum Mutual Information Feature Extraction Method Based on the Cloud Platform

Wei Shasha¹, Lu Huijuan¹, Jin Wei¹, Li Chao²

(1. College of Information Engineering, China Jiliang University, Hangzhou 310018, China;

2. College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou 310018, China)

Abstract: With the large-scale application of gene chip, gene expression data with high dimension which exists a large number of irrelevant and redundant features may reduce classifier performance problem. A maximum mutual information feature extraction method based on cloud platforms was proposed. Hadoop cloud computing platform could be a parallel computing after gene expression data segmentation, features was extracted at the same time combined with the maximum mutual information method and the characteristics of cloud computing platform filter model was realized. Simulation experiments show that the maximum mutual information feature extraction method based on the cloud platform can rapid extraction of features in a higher classification accuracy which save a lot of time resources to make a highly efficient gene feature extraction system.

Key words: cloud computing, maximum mutual information, cloud platform, feature extraction

1 引言

基因芯片技术是随着人类基因组计划发展起来的一项新技术, 大规模基因芯片(DNA 微阵列)技术的应用是现在功能基因组以及肿瘤诊断等研究中的重要监测手段^[1], 可广泛用于基因序列分析、基因突变检测、

疾病诊断等诸多领域。基于基因表达数据维数高、样本小的特点, 直接对其进行学习的生物学分析成本较高, 并且有些基因只在特定实验条件下表达, 为了降低机器学习的时间及空间复杂度, 需要对其进行特征选择, 选取与分类紧密关联的基因, 同时提高分类精度^[2]。特征选择是根据各个特征的重要程度, 剔除特征中不相

^{*} 国家自然科学基金资助项目(No.61272315, No.60842009, No.60905034), 浙江省自然科学基金资助项目(No.Y1110342, No.Y1080950)

关的冗余特征后,挑选出对分类有意义的某些特征,以降低特征空间维数。在模式识别、数据挖掘以及机器学习,特征选择都非常关键^[3]。特征选择方法从研究之初到现在,已经有了很多成熟的方法,2005年,Peng^[4]和 Ding^[5]在处理连续特征时,分别使用 F-Statistic 和 Pearson 相关系数度量相关性程度进行特征选择。这为本文利用两个特征之间互信息最大化的方法进行特征过滤提供了思路与基础。

互信息是信息论中的一个重要概念,通常用于描述两个随机变量间的统计相关性,用一个变量中包含另一个变量的信息多少表示两个随机变量之间的依赖程度,一般用熵表示^[6]。在统计学上,同一分类系统的基因并非独立而是相关的,确定基因之间的互信息就是要定义相似性测度寻找变换关系,使得基因间的相似性达到最大,从而确定该基因在分类系统中的重要程度。其中,当信息量达到最佳配准时,即实现互信息最大化。在医学领域中,利用互信息最大化法进行多模医学图像配准包括 CT 扫描以及核磁共振等成为医学图像处理方面的热点^[7],它能够很快地排除很大数量的非关键性的噪声和无关基因,是特征选择中一种基于相关性的过滤方法。但在数据量较大的情况下,对服务器性能要求高,计算效率低,而云计算的出现为解决这个问题提供了新的契机^[8]。

云计算是传统计算机技术发展融合的产物,是一种基于互联网的计算方式,通过这种方式,共享的软硬件资源和信息可以按需提供给计算机和其他设备^[9]。云是网络、互联网的一种比喻说法,一般来说包括 IaaS、PaaS 和 SaaS 3 个层次。

目前,将云计算与分类问题结合,参考文献[8]等提出了基于云计算平台的代价敏感集成学习算法,参考文献[10]提出了云计算在贝叶斯分类中的应用,但还没有相关文献讨论如何在云计算平台环境下对数据特征进行过滤。为了能够快速、准确、高效地处理基因数据特征提取问题,本文提出了一种基于云计算平台的 Filter 型基因选择算法——CMI-Selection。

本文对算法的实现进行了仿真模拟,在 CMI-Selection 中,实验室用 5 台 PC 搭建了 Hadoop 云计算平台,首先利用随机函数将数据集随机分为 5 个部分,每个部分遵循互信息算法进行筛选计算,然后将结果返回到客户端用于测试及分类。为了评估算法在云平台下的性能,实验用 ELM(极限学习机)对提取后的特征进行训练和学习,结果表明,在分类精度与普通 PC 端相近的情况下,CMI-Selection

速度更快。

2 互信息与基因筛选

在进行基因选择和基因降维之前,首先要进行基因筛选。在基因表达数据特征提取的前期过程中,基因筛选能够提高计算效率,是选出具有代表性的精简的基因子集的有效方法^[11]。

熵用来表示任何一种能量在空间中分布的均匀程度,其大小跟能量分布均匀程度有关,能量越不确定且分布越均匀,熵就越大^[12]。信息论中的“信息熵”是香农^[13]在进行信息处理时提出的概念。在互信息最大化方法中,信息熵主要用来衡量一个随机变量取值的重要性程度,以特征能够为分类带来多少信息为衡量标准,带来的信息越多,该特征越重要。

假设 X 来自于一个集合 S , 且 X 是一个离散随机变量。 X 的概率密度分布函数表示为 $p(x)$, 则 X 的信息熵定义如下:

$$H(X) = - \sum_{x \in S} P(x) \lg(P(x)) \quad (1)$$

已知一个来自于集合 T 的变量 Y , $P(x|y)$ 表示 Y 取值为 y 、 X 取值为 x 的概率, X 的不确定性用 $H(X|Y)$ 衡量, 如式(2)所示:

$$H(X|Y) = - \sum_{y \in T} P(y) \sum_{x \in S} P(x|y) \lg(P(x|y)) \quad (2)$$

$P(x,y)$ 用来表示 X 、 Y 的联合概率密度, 则它们的互信息 $I(X;Y)$ 定义为:

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x \in S} \sum_{y \in T} P(x,y) \lg \frac{p(x,y)}{p(x)p(y)} \quad (3)$$

在基因筛选过程中, 互信息通常用来表示计算特征与特征之间的关系。在上述式子中, 假设考虑特征 t 与类 c 的分布, N 为基因总数, A 为类 c 中出现特征 t 的基因数, B 为非类 c 中出现特征 t 的基因数, C 为类 c 中不出现特征 t 的基因数, 特征 t 与类 c 之间的互信息定义为:

$$I(t;c) = \lg \frac{p(t|c)}{p(t)} = \lg \frac{p(t,c)}{p(t) \times p(c)} \approx \lg \frac{A \times N}{(A+C) \times (A+B)} \quad (4)$$

如果 $I(t;c)=0$, 那么特征 t 与类 c 相互独立。

在式(5)中提供关于类别信息的加权平均值来衡量一个特征在全局特征选择中的重要性:

$$I_{\text{avg}}(t) = \sum_{i \in 1}^k P(c_i) I(t, c) \quad (5)$$

特征选择后, 尽可能多地保留关于类别的信息, 即达



到互信息最大化:

$$\max MI(t) = \sum_{i=1}^k P(c_i|t) \lg \frac{P(c_i|t)}{P(c_i)} \quad (6)$$

则最大互信息量为:

$$I^* = \max MI(t) \quad (7)$$

用互信息最大化方法进行特征选择后,选出来的特征集合应该尽可能多地提供关于某个类别的信息。两个随机变量之间共有的信息量越大,则两个变量之间的相关程度越高,互信息量越大;如果两个随机变量完全不相关,则两个变量之间不相关,互信息量为 0。互信息最大化是选择相关程度最高的两个变量进行循环迭代,得出每次信息量中相关程度最高的变量。

3 基于云计算平台的特征选择系统

云计算^[4]描述了一种基于互联网的新的 IT 服务增加、使用和交付模式,通常涉及互联网提供动态易扩展而且经常是虚拟化的资源,意味着计算能力也可以作为一种商品进行流通。云是网络、互联网的一种比喻说法。典型的云计算提供商往往提供通用的网络业务应用,软件和数据都存储在服务器上(即云端),用户可以通过浏览器或者其他 Web 服务访问。

3.1 Hadoop 云计算平台的体系结构

Hadoop 是 Lucene 子项目 Nutch 的一部分,是由 Apache Software Foundation 开源组织提出的一个分布式计算开源框架,它不是一个缩写字,而是一个虚构的名字。Hadoop 的核心是 MapReduce 和分布式文件系统 (Hadoop distributed file system, HDFS) 及后来加入的 HBase。MapReduce 就是任务的分解与汇总(规约),如图 1 所示,它是一种简化的并行计算编程模型,其中 map 主要是把任务分解成多个任务,而

reduce 则把分解后的多任务处理汇总起来。

HDFS 用来存储分布式计算的数据,采用 master/slave 架构,由若干个数据节点和一个名称节点组成。服务器间相互通信的过程如图 2 所示。

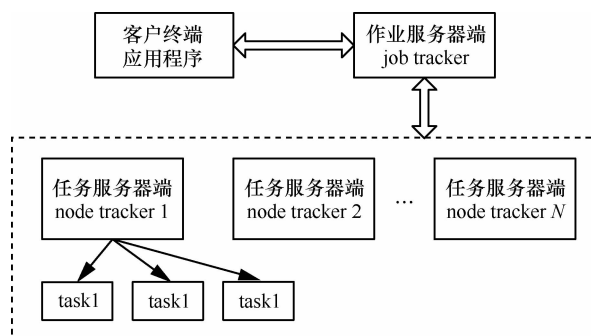


图 2 服务器之间相互通信的过程

HBase 则对应于 Google 的 BigTable。

Hadoop 的特点介绍如下。

- 构建成本低: Hadoop 框架对硬件环境没有任何限制,无需昂贵的服务器,普通的 PC 即可实现。在软件使用方面,由于部署目标平台 Linux 是开源的,不存在软件授权费等方面的问题。
- 可靠性: 在实现时,由于 Hadoop 认为所有节点都有可能发生计算或者存储失败,故其在节点群中维护了很多工作副本,一定程度上保证了系统的备份恢复机制和分布式处理的可靠性。
- 扩容能力: 能非常容易地增添计算存储资源。
- 效率高。

3.2 基于 MapReduce 的特征过滤

MapReduce 由两个动词组成,分别控制任务的分解和汇总,从技术创新角度来讲,MapReduce 也并不是创新技术,分布式并行计算程序的编写也十分简单, Hadoop 中 Streaming 工具使用起来方便快捷,一般编程技术就可以开发出一个分布式并程序,用于海量数据的并行计算。

在对基因表达数据进行特征过滤中,步骤如下。

(1) 随机函数对基因数据集进行随机分块。为了仿真模拟,实验用 5 台 PC 组成一个 Hadoop 云计算平台。

(2) map 函数对分块的特征集进行信息熵的计算。这里的 map 函数被定义为互信息最大化算法。通过设置 map 任务的特征数量的大小,让云平台自动对 t 时刻到达的特征集进行划分,每块数据对应一个 map 任务,每个 map 任务计算各自特征集的信息熵,同一时刻不同 map 之间进行并行计算,得到 t 时刻所有特征的信息熵。

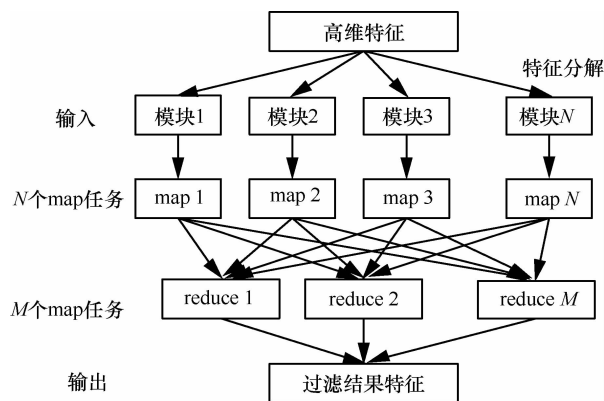


图 1 MapReduce 工作过程

(3)执行 reduce 任务,包括特征提取和特征集成两个阶段。其中特征提取按照互信息最大化算法的标准进行。

4 实验结果

在理论分析的基础上,本节选取 4 组基因表达数据集对 CMI-Selection 进行性能测试,其中 Breast、Colon、Heart 为两类数据,Leukemia 为多类数据。数据集信息见表 1。

表 1 数据集信息

数据集	样本总数	基因数	类分布	
			类名	样本数
Breast	97	24 481	relapese	46
			non-relapese	51
Leukemia	72	7 129	ALL	24
			MLL	20
			AML	28
Colon	62	2 000	negative	40
			positive	22
Heart	270	3 510	negtive	150
			positive	120

在特征提取之前,先将基因表达矩阵中的元素进行对数转换,实现标准化:

$$x_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{\frac{1}{N-1} \sum_{j=1}^N (x_{ij} - \bar{x}_i)^2}} \tag{8}$$

本文用 Breast 数据集作为实验例子进行分析与研究,首先用随机函数对数据集进行分块,在云平台上 4 台 PC 部署数据节点和任务服务器端,即每部分得到 6 120 个特征。每块特征数据对应一个 map 任务,每个 map 任务计算各自特征集的信息熵,利用互信息最大化方法得出特征集额互信息,随后开始执行 reduce 步骤。在 reduce 步骤中,对上一步得到的互信息进行排序,筛选特征,得到排名前 1 224 个特征。最后进行汇总,运送到客户机,在客户机端用 ELM 对获得的基因特征进行训练和测试。实验结果如图 3、图 4 所示。

从图 3 可以看出,在云平台环境下,通过互信息对特征进行筛选后,将特征和标签分别作为 ELM 的输入和输出进行训练与测试,采用 5 折交叉验证,获得的精度最高可以达到 93%,与在普通 PC 环境中相同特征数量的前提下进行比较可以发现其分类精度大致相同,说明 CMI-Selection 算法在分类精度方面具有可行性与有效性,它能够保证提取的特征是有效的,具有较高的分类精度。

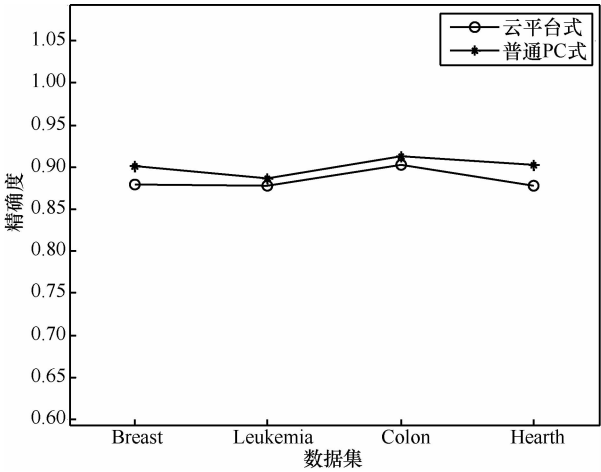


图 3 ELM 在不同环境下的不同数据集上的分类精度

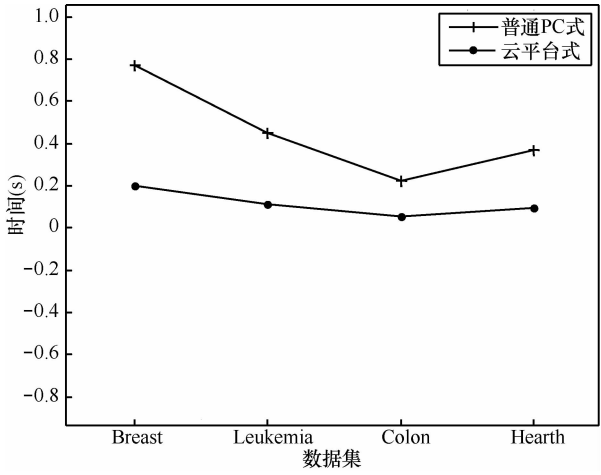


图 4 不同环境下的特征基因提取时间

从图 4 中可以看出,由于云平台的并行计算性能,其在保证较高分类精度的同时,相比于普通 PC 速度提高了 4 倍左右,并且这种提速会随着服务器数量的增加变得更加明显,从而为大数据学习节省了时间资源,说明了云平台的高度并行化。

5 结束语

本文将传统的特征提取方法同云计算平台相结合,提出了一种基于云平台的互信息最大化的特征提取方法,构建了一个基于云平台的特征过滤模型。该模型通过设置平台中 map 任务的个数、划分特征集的大小让云平台自动对基因数据集进行特征的筛选提取,提取后的特征用于训练与测试。与普通单个 PC 相比,该模型在保证较高的分类精度的情况下,速度快,易于实现。实验结果表明,基于云平台的互信息最大化特征提取方法是正确、可行的,能够快速提取特征,节省大量时间资源,是一种高效的基因特征



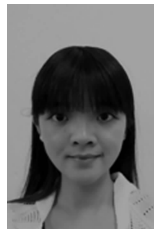
提取系统。

参考文献

- 1 Kang H N, Chen I M, Wilson C S. Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. *Blood*, 2010(115):1394~1405
- 2 任江涛, 黄焕宇, 孙婧昊. 基于相关性分析及遗传算法的高维数据特征选择. *计算机应用*, 2006, 26(6): 1403~1405
- 3 裴国永, 王娜, 汪万紫. 基于互信息和遗传算法的两阶段特征选择方法. *计算机应用研究*, 2012, 29(8): 2903~2905
- 4 Peng H H, Long F H, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(8): 1226~1238
- 5 Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *Journals of Bioinformatics and Computational Biology*, 2005, 3(2): 185~205
- 6 王凌, 陈震, 危水根等. 基于改进最大互信息法的 MR 切片图像配准. *生物医学工程学杂志*, 2012, 29(2): 201~205
- 7 杨虎, 马斌荣, 任海萍等. 基于最大互信息的人脑 MR-PET 图像配准方法. *北京生物医学工程*, 2001, 20(4): 246~251
- 8 张铃卫, 王文强. 基于云计算平台的代价敏感集成学习算法研究. *山东大学学报(工学版)*, 2012, 42(4): 19~23
- 9 Vouk M A. Cloud computing-issues, research and implementations. *Proceedings of ITI 2008, Dubrovnik*, 2008: 79~120
- 10 朱杰. 云计算在基于贝叶斯分类的垃圾短信过滤中的研究与应用. *电子科技大学硕士学位论文*, 2010

- 11 王明怡. 微阵列数据挖掘技术的研究. *浙江大学博士学位论文*, 2004
- 12 刘庆和, 梁正友. 一种基于信息增益的特征优化选择方法. *计算机工程与应用*, 2011, 47(12)
- 13 Hu Y, Loizou P C. Speech enhancement based on wavelet thresholding the multitaper spectrum. *IEEE Transactions on Speech and Audio Processing*, 2004, 12(1): 59~67
- 14 戴元顺. 云计算技术简述. *信息通信技术*, 2010(2)

[作者简介]



魏莎莎, 女, 中国计量学院信息工程学院硕士研究生, 主要研究方向为云计算、机器学习、数据挖掘。

陆慧娟, 女, 博士, 中国计量学院信息工程学院教授, CCF 常务理事、高级会员, 主要研究方向为云计算、机器学习、模式识别和生物信息学等。

金伟, 男, 中国计量学院信息工程学院硕士研究生, 主要研究方向为模式识别、图像处理。

李超, 男, 中国计量学院机电工程学院硕士研究生, 主要研究方向为云计算等。

(收稿日期: 2013-09-16)

(上接第 30 页)

5 结束语

目前电信运营商的云数据中心业务投资效益还有提升空间。可以预见, 电信运营商将会发展各种云增值业务, 利用现有资产进行建设, 形成规模效应, 不断降低成本, 提高项目效益, 推动国内云数据中心业务的蓬勃发展。

电信运营商推出云数据中心业务, 对云计算领域的竞争格局造成了较大影响, 其他云计算服务提供商需要了解电信运营商的业务发展情况和策略, 以制定自己的应对措施。本文以假设场景和典型项目进行实证研究, 其他场景和项目也可以参考本研究的模型进行分析。

参考文献

- 1 国家发展和改革委员会, 建设部. 建设项目经济评价方法与参数(第 3 版). 北京: 中国计划出版社, 2006

- 2 何明, 郑翔, 赖海光等. 云计算技术发展及应用探讨. *电信科学*, 2010, 26(5)

[作者简介]



冯若钊, 男, 中国电信股份有限公司广东研究院高级经济师, 主要从事投资规划和投资管理、成本研究和项目效益评估、网络规划管理和平台开发等方面的研究工作。

范云凌, 女, 中国电信股份有限公司广东研究院高级工程师, 主要从事通信市场规划、网络规划、投资规划以及项目和业务经济评估方法、工具设计等方面的研究工作。

(收稿日期: 2013-09-01)