第29卷, 第4期
2009年4月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 29, No. 4, pp1074-1077
April, 2009

分布式并行计算在光谱学信号处理领域中的应用

陈永明, 林 萍, 鲍 丹, 何 勇*

浙江大学生物系统工程与食品科学学院, 浙江 杭州 310029

摘 要 将分布式并行计算引入光谱学信号处理领域。用傅里叶红外光谱仪 FTIR-4100 获得白砂糖、木糖醇、麦芽糖和葡萄糖4类糖各39个样本的光谱曲线作为测试数据。在两台软硬件配置相同的计算机平台上运行分布式并行算法。先运用分布式并行方法读取 FTIR-4100 生成的文本文件中的原始数据, 然后进行分布式并行数据预处理, 包括最大峰值标准化校正, Savitzky-Golay 平滑降噪算法等, 再运用分布式并行遗传算法抽取糖特征波数共24个, 最后将提取到的24个特征波数作为用 BP 神经网络输入, 建立3层人工神经网络。实验结果表明, 分布式并行计算运行结果与单机顺序计算结果比对一致, 在两台计算机并行工作模式下的计算效率比传统的单机顺序计算处理效率高33.6%, 为光谱学信号处理研究领域进行复杂科学计算和提高计算效率提供了新的方法。

关键词 分布式并行计算; 信号处理; 糖; 复杂科学计算; 计算效率
中图分类号: TP338.8 文献标识码: A **DOI**: 10.3964/j.issn.1000-0593(2009)04-1074-04

引 言

在光谱学信号处理领域中研究人员多数关心信号处理的方法, 而较少关心信号处理的效率, 例如在光谱学相关的期刊和相关的著作[1, 2]中分布式并行算法应用鲜有论述。然而随着传感器技术和计算机技术不断进步, 光谱仪器精密化程度不断提高, 测量范围和分析领域不断扩大, 测量样品的复杂性和数量不断增加, 传统的顺序信号处理模式越来越暴露出其处理速度低下的弊端, 大批的数据往往要处理好几天, 大大影响了研究人员的工作效率。有些需光谱实时处理和分析信息的场合, 更是需要能快速完成数据处理。因此, 如何采用先进的计算机技术、数学算法和现有的资源来提高研究效率即显得尤为重要。国内外相关的研究机构已经都注意到并着手解决这个问题, 例如 IBM 公司开发用于人机国际象棋大战的 Deep Blue 计算机, 北京大学开发研究的湍流并行计算机等即是用来提高计算效率、满足复杂科学运算的需求。

然而, 专业的高端并行计算机购买与开发成本高, 一般的实验室都无法利用现有的高性能计算机技术。因此, 如何利用现有的计算机资源来实现高速计算才是可行之路。针对这个问题, 本文提出了基于 Matlab Ver7.5.0 的分布式并行计算工具(distributed computing toolbox, DCT)并结合个人电脑和互联网技术完成分布式并行计算方法。

1 材料与方法

1.1 仪器设备和处理软件

实验使用日本 JASCO 公司的傅里叶红外光谱仪 FTIR-4100 获得红外光谱数据, 其测定范围在 349~7 801 cm^{-1}之间, 设定分辨率为 1 cm^{-1}。分布式并行数据处理使用两台计算机硬件平台均为 Pentium 4 CPU 3.00 GHz, 512 MB DDR 内存。软件平台也均使用 Windows XP 操作系统与数学分析软件 MATLAB Ver7.5.0。

1.2 样品来源及光谱的获取

从超市买来四种糖分别是太古纯正白砂糖(Sucrose)、禾甘木糖醇(Xylitol)、双歧麦芽糖(Maltose)、红芍口服葡萄糖(Dextrose)。各取1包样本, 每包400 g。每种糖与溴化钾(KBr)按1∶19的比例均匀混合, 用压片机压实成直径为5 mm, 厚度为2 mm 的圆柱体, 放入已标定好的 FTIR-4100 光谱仪做透射实验。每个品种各做39个样本, 共计156个样本。图1为随机选取四种不同糖各3个样本红外光谱曲线图。

1.3 光谱数据预处理

收稿日期: 2007-12-21, 修订日期: 2008-03-25
基金项目: 国家"十一五"支撑计划项目(2006BAD10A0403), "863"计划项目(2007AA10Z210), 浙江省自然科学基金项目(Y307158)和浙江省教育厅项目(20071064)资助
作者简介: 陈永明, 1982年生, 浙江大学生物系统工程与食品科学学院博士生 * 通讯联系人 e-mail: yhe@zju.edu.cn

重复测量光谱样品实验中，由于受样品的颗粒度和样品间散射等因素的影响，会引起的光谱曲线漂移现象，采用最大峰值标准化校正来降低光谱曲线基线漂移的影响。实际测量光谱曲线经常混有不同程度的噪声，为了降噪又必须保留峰值的形状或者其包含的一些高频成分，实验采用 Savitzky-Golay 算法进行降噪，在很大程度上保证了测量光谱峰值的有效性。

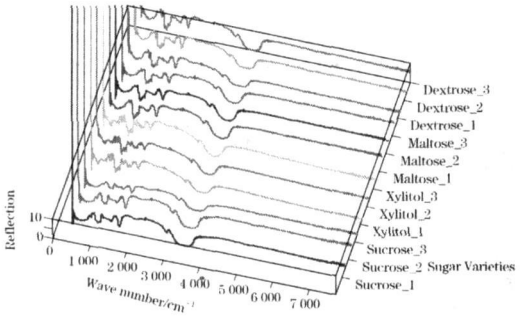

Fig 1　Infrared reflectance spectra of 4 different varieties of sugar

1.4　遗传算法选择波长

遗传算法是模拟生物进化机制进行随机优化的一种算法，将其应用于波长选择，其主要步骤有：染色体编码、种群初始化、适应度函数、遗传操作、设定停止条件和波长选择[3, 4]。

1.5　BP 神经网络

BP 人工神经网络技术具有较强的人工智能功能和模拟多元非线性体系的能力，与传统的线性回归技术相比，它不但具有自适应和自组织能力，而且它的突出优点在于其强大的非线性映射能力，因此很方便地用其来预测样本的种类[5, 6]。

1.6　Matlab 分布式并行计算

分布式并行计算是把原来由单台计算机独立完成任务分配给网络上其他计算机协同完成计算的一种算法[7-10]。分配任务的计算机叫客户端(Client)，被分配到任务的计算机叫做服务器(Service)。网络上每台计算机都可以拥有“监工”(Job Manger)，它主要完成顺序工作和分配任务给工人(Worker)，工人主要是做被分配给的任务计算。图 2 为 Matlab 分布式并行工作管理模式。Matlab 提供分布式并行计算法的工具箱，用于实现分布式并行计算。

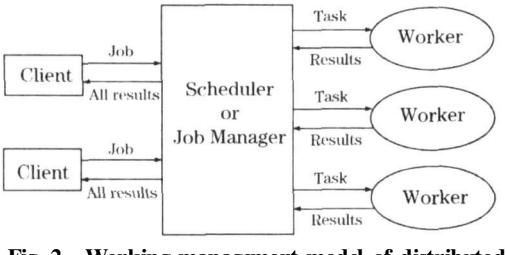

Fig 2　Working management model of distributed computing in matlab

用 Matlab 实现并行计算主要步骤如下。

寻找网络上可用的 Job Managers 和 Workers。

创建由一系列任务组成的工作。

创建由 Worker 单独完成的计算任务。

分配相应任务给各个 Worker。

Job Manager 取得结果。

2　试验结果与分析

2.1　光谱数据预处理分布式并行实现

实验中需要处理的光谱数据量大于 $2\ 411\times 10^6$，采用传统单机顺序处理方法，无疑将耗费大量时间。将这些需要处理的数据放到 Windows 映射网络驱动器上，这样网络上的计算机就可共享数据，每个 Worker 读取 Job Manager 分配的原始数据，各自完成标准化校正，信号降噪，从而实现分布式并行数据预处理。

2.2　分布式并行遗传算法抽取特征波长分析

FTIR-4100 光谱仪测量波长范围为范围 349～7 801 cm^{-1}，共有 7 729 个波数。若将这 7 729 个波数直接作为 BP 神经网络的输入变量，不但运算量大，而且当光谱曲线特征差异不明显时，它们无法将不相关性或非线性变量剔除，即无法建立正确的校正模型。实验中用遗传算法分别抽取蔗糖与木糖醇、木糖醇与麦芽糖、麦芽糖与葡萄糖两两之间各 8 个特征波数，共 24 个特征波数作为 BP 神经网络的输入变量。由此可见，使用遗传算法有效地进行了数据压缩，为将来 BP 神经网络预测提供了更强的校正模型。

实验中遗传算法每次只抽取 2 类不同的糖特征波数，将不同类型糖进行两两排列后由 Job Manager 分别分配给不同的计算机上的 Workers 进行遗传算法特征波数的提取，最终将提取的波数反馈给 Job Manager 实现分布式并行遗传算法。图 3 虚线对应的横坐标即为用分布式并行遗传算法提取的 Xylitol 与 Maltose 可鉴别特征波数。

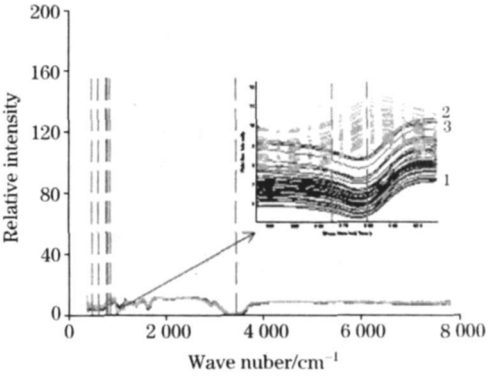

Fig 3　Spectrometry data with discriminative features found by GA

1: Xylitol; 2: Maltose; 3: Discriminative features

2.3　分布式 BP 神经网络预测模型

利用遗传算法抽取的 24 个特征波数作为输入，中间层设定为 5，糖品种类型作为神经网络输出(将蔗糖、木糖醇、

麦芽糖、葡萄糖品种类别分别设置为 1, 2, 3, 4)。建立一个 4 层输入单元, 5 个隐含单元和 1 个输出单元的 BP 神经网络。取 4 种不同糖品种各 30 个样本, 共 120 个作为建模样本, 对未知样本 36 个进行预测。预测结果表明, 建模样本品种类型的拟合率和预测识别率均为 100%, 预测结果见表 1。

Table 1 Discrimination results of unknown 36 samples of sugar

序号	类别	预测值	序号	类别	预测值
(1)	1	1.021	(19)	3	3.001
(2)	1	0.999	(20)	3	3.001
(3)	1	1.002	(21)	3	2.989
(4)	1	1.003	(22)	3	2.999
(5)	1	1.003	(23)	3	2.999
(6)	1	1.002	(24)	3	2.998
(7)	1	1.003	(25)	3	2.999
(8)	1	1.006	(26)	3	3.001
(9)	1	1.002	(27)	3	2.994
(10)	2	2.016	(28)	4	3.995
(11)	2	2.001	(29)	4	3.992
(12)	2	2.003	(30)	4	3.989
(13)	2	1.999	(31)	4	3.991
(14)	2	2.000	(32)	4	3.997
(15)	2	2.000	(33)	4	3.996
(16)	2	2.000	(34)	4	3.999
(17)	2	2.000	(35)	4	3.991
(18)	2	2.000	(36)	4	3.998

Note: 类别 1~ 4 分别代表蔗糖、木糖醇、麦芽糖和葡萄糖

2.4 分布式并行计算结果

分布式并行算法是将传统一台计算机顺序完成的任务分配给网络上其他计算机系统协同完成的算法, 因此相对顺序计算方法是一种高效的计算方法。本实验分布式并行算法在两台软硬配置相同的计算机上进行测试, 完成从 FT-IR-4100 光谱仪导出的糖文本文件的数据的读取, 最大峰值标准化校正, Savitzky-Golay 降噪, 遗传算法抽取特征波数, BP 神经网络预测等一系列分布式并行计算。表 2 为顺序计算与并行计算耗时对比。实验结果表明, 在两台计算机上运行分布式并行算法比在其中一台计算机运行顺序计算速度效率高 33.6%。

Table 2 Consumption time of the serial and parallel computation

运算模式	读文件与预处理/s	遗传算法/s	神经网络/s	总计/s
顺序计算	82.329	62.422	1.173	145.924
并行计算	67.430	28.590	0.914	96.934

3 结 论

提出了分布式并行计算在光谱学信号处理领域中的应用。将糖的红外光谱数据运用分布式并行算法进行测试与研究。实验结果表明, 在两台软硬配置相同的计算机上运行并行计算比在其中一台计算机进行顺序运算速度效率高 33.6%。如果增加网络上计算机的数量, 无疑将大大提高程序处理的运行速度, 从而提高研究人员的工作效率与积极性。分布式并行算法是将传统一台计算机完成的任务分配给网络上其他计算机系统协同完成的高效算法。因此, 其为研究人员将来光谱学信号处理研究中遇到复杂的科学计算和提高研究效率提供有效的解决方法。由于提出的分布式并行计算方法是基于 Matlab Ver7.5.0 的分布式并行工具箱并结合当今个人电脑和互联网技术下完成运算的, 因此研究方法具有很强的可行性与实用性。

参 考 文 献

[1] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, et al(陆婉珍, 袁洪福, 徐广通, 等). Modern Near Infrared Spectroscopy Analytical Techonlogy(现代近红外光谱分析技术). Beijing: China Petrochemical Press(北京: 中国石油化工出版社), 2007. 33.

[2] YANG Yan-lu, ZHAO Long-lian, HAN Dong-hai, et al(严衍禄, 赵龙莲, 韩东海, 等). The Application and Foundation of NIR Analysis(近红外光谱分析基础与应用). Beijing: China Light Industry Press(北京: 中国轻工出版社), 2005. 98.

[3] CHENG Biao, WU Xiao-hua, CHEN De-zhao(成 飙, 吴晓华, 陈德钊). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(10): 1923.

[4] LIU Fang, WANG Jun-de(刘 芳, 王俊德). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2002, 22(2): 239.

[5] HUANG Min, HE Yong, CEN Hai-yan, et al(黄 敏, 何 勇, 岑海燕, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27(5): 916.

[6] HE Yong, LI Xiao-li, SHAO Yong-ni(何 勇, 李晓丽, 邵咏妮). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(5): 850.

[7] SHI Jia-ru, ZHENG Shu-xin, CHEN Huai-bi, et al(施嘉儒, 郑曙昕, 陈怀璧, 等). High Energy Physics and Nuclear Physics(高能物理与核物理), 2005, 29(8): 818.

[8] ZHAO Jun, SONG Jun-qiang, LI Zhen-jun. Advances in Atmospheric Sciences, 2003, 20(1): 159.

[9] GAO Tian-chi, LI Yue-lian(高天池, 李月莲). Journal of Shanghai Jiaotong University(上海交通大学学报), 2005, 39(6): 979.

[10] HU Jun, GUO Shao-zhong, ZHOU Bei(胡 军, 郭绍忠, 周 蓓). Computer Engineering(计算机工程), 2007, 33(5): 68.

Application of the Distributed and Parallel Computation in Spectroscopy Signal Processing

CHEN Yong-ming, LIN Ping, BAO Yi-dan, HE Yong*
College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China

Abstract The distributed and parallel computation was introduced to spectroscopy signal processing. The reflection spectra of 4 different varieties of sugar including sucrose, xylitol, maltose and dextrose were measured with FI/IR-4100 Fourier infrared spectral equipment. Each type of sugar consisted of 39 samples. The distributed and parallel algorithm was executed on 2 computers with the same hardware and software systems. First, the distributed and parallel algorithm was used to read original spectral data from the text files generated by FT/IR-4100 device. Second, the data were preprocessed by distributed and parallel algorithm. The preprocessing methods include standard normalization to the maximum peak, Savizky-Golay smoothing denoising, etc. Third, search for the key discriminative wave numbers in mass spectrometry data was performed by distributed and parallel genetic algorithm (GA). At the end, the discriminative features of 24 wave numbers extracted by GA were applied as BP neural network inputs and a 3-layer neural network was built up. The computing results generated by distributed and parallel algorithm are the same as the serial computing results generated by single personal computer. The processing efficiency using 2 personal computers is 33.6% higher than that of serial computation. So the paper presents a creative method for the complex scientific computation and enhancing the computing efficiency in spectroscopy signal processing.

Keywords Distributed and parallel computation; Signal processing; Sugar; Complex scientific computation; Computing efficiency

(Received Dec. 21, 2007; accepted Mar. 25, 2008)

* Corresponding author