

# 面向多源异构信息的频繁项集挖掘算法

刘自力<sup>1</sup>, 范军丽<sup>2</sup>, 陈文伟<sup>3</sup>, 吴润泽<sup>3</sup>

(1. 国网山西省电力公司 晋城供电公司, 山西 晋城 048000;

2. 北京国电网络技术有限公司, 北京 100070;

3. 华北电力大学 电气与电子工程学院, 北京 102206)

**摘要:** 电网调度运行过程中产生海量复杂度高的多源异构数据, 利用数据挖掘将这些数据转化为知识是调度智能化发展的必然趋势。为此, 构建了基于调控大数据的多源异构数据分析模型, 提出了一种能够处理大数据的频繁项集挖掘算法, 将分布式统计引入到频繁项集挖掘过程。该算法根据组合学原理, 利用 MapReduce 扫描一次数据库从原始事务数据库中完成频繁项集的整个挖掘过程; 且在支持度阈值改变的情况下无需重新扫描数据库进行挖掘, 改进了现有频繁项集挖掘算法多次扫描事务数据库和挖掘效率低的问题。通过利用 Hadoop 平台对故障信息事务库进行处理, 将所提出的算法与其他频繁项集挖掘算法进行了对比验证实验。实验结果表明, 所提出的算法不受支持度阈值的影响, 处理海量事务数据算法时间开销小, 可为实现以准确、安全、经济等目标综合最优的调度智能化分析和决策提供有益的知识。

**关键词:** 智能调度; 频繁项集; 组合理论; Hadoop

**中图分类号:** TP39

**文献标识码:** A

**文章编号:** 1673-629X(2017)06-0076-05

**doi:** 10.3969/j.issn.1673-629X.2017.06.016

## Frequent Itemset Mining Algorithm for Multi-source Heterogeneous Information

LIU Zi-li<sup>1</sup>, FAN Jun-li<sup>2</sup>, CHEN Wen-wei<sup>3</sup>, WU Run-ze<sup>3</sup>

(1. Jincheng Power Supply Company, State Grid Shanxi Electric Power Company, Jincheng 048000, China;

2. Beijing Guodiantong Network Technology Co., Ltd., Beijing 100070, China;

3. School of Electrical and Electronic Engineering, North China Electric Power University, Beijing 102206, China)

**Abstract:** Power grid dispatching has produced large amount of multi-source heterogeneous data with high complexity, and it is the inevitable development trend of intelligent dispatching that power data are transformed into knowledge by data mining. An analysis model of multi-source heterogeneous data based on big data in power dispatching and control system has been established and a frequent item set mining algorithm for processing big data has been proposed. The distributed statistics has been introduced into mining frequent item sets. Combining MapReduce programming and combinatorics, the target frequent item set mining has been completed via scanning transaction database with the proposed algorithm and thus there is no need to scan database again for mining while support degree is under variation. This algorithm has been promoted to solve the problem of multiple scanning transaction database and low mining efficiency. Compared with other frequent item set mining, the algorithm takes advantage of Hadoop in dealing with fault information transaction database. Experimental results show that the proposed algorithm performs well in expansibility and has less time cost with large transaction database and that the method adopted has provided useful knowledge for intelligent analysis and decision making with comprehensive optimal objectives of accuracy, security, economic and others, which single data source could not achieve.

**Key words:** intelligent dispatching; frequent itemsets; combinatorics; Hadoop

## 0 引言

当前信息通信技术(ICT)的高速发展推动了智能

电网的全面建设, ICT 和电网建设的深度融合催生了智能电网大数据的爆炸性增长。这些数据不仅规模

收稿日期: 2016-06-20

修回日期: 2016-09-22

网络出版时间: 2017-04-28

基金项目: 国家自然科学基金资助项目(51507063)

作者简介: 刘自力(1969-), 男, 电力高级工程师, 研究方向为电力通信技术及网络规划。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20170428.1702.028.html>

大,其结构也多样化,构成了智能电网中的海量多源异构大数据。海量多源异构大数据的高效快速处理和深度挖掘分析为建设坚强可靠、稳定运行的智能电网提供基础<sup>[1]</sup>。电网调度控制系统中产生的大数据是智能电网大数据的主要来源,这些数据隐藏着电网运行中的实时状态信息。而数据挖掘是实现将实时数据和沉淀的历史数据转化为有用知识的有效方法,为电网调度运行提供辅助性决策和科学性建议<sup>[2]</sup>。智能调度中的数据来源丰富,人们从不同的数据源获取的信息也越来越多,而且这些数据创造的价值也被人们所接受,但是这些数据源之间形成了众多的“信息孤岛”。因此,有必要采用大数据思想,对智能调度中的数据进行分析和挖掘,来实现数据共享,为智能调度的实现提供参考。关联分析是数据挖掘和知识发现的重要技术之一。关联规则算法主要用来挖掘事务数据库中有意义或用户感兴趣的规则。1993年,美国的 R. Agrawal 等首次提出关联规则算法<sup>[3]</sup>,其主要思想是从原始事务数据库中找出满足一定支持度和置信度要求的项集。其中满足用户定义的最小支持度的项集称为频繁项集,从频繁项集中找出满足最小置信度的频繁项集,将其转化成最终的强关联规则形式完成关联规则挖掘。将频繁项集转化成关联规则的过程较为简单,所以频繁项集挖掘是关联规则挖掘的重点和关键。

Apriori 算法是由 R. Agrawal 等提出的经典频繁项集算法,该算法通过连接和剪枝方法完成,能够有效挖掘出用户需要的关联规则,但是存在产生大量的候选项集和重复多次扫描事务数据库的缺陷。为了克服这些缺陷,Han Jiawei 等在 Apriori 算法的基础上提出了 FP-growth 算法<sup>[4]</sup>。该算法建立了树结构,用来保存每项的支持度计数,在建立频繁模式树的过程中只需扫描两次事务数据库,并且不产生候选项集。尽管该算法提高了频繁项集的挖掘效率,但是针对大量且事务比较长的事务数据库,其挖掘效率较低。为了提高频繁项集的挖掘效率,文献[5-8]基于磁盘存储的算法改进了挖掘大量事务数据库和内存有限的问题,但其算法复杂度较高。文献[9]的 FIMM 算法在挖掘频繁项集的过程中其运行时间不受支持度阈值的影响,改善了算法的计算复杂度;然而,事务数据库数据量很大时,其结果也不理想。

针对以上问题,在分析智能调度中数据特点的基础上,建立了智能调度多源异构数据分析模型,实现了多源异构数据为智能调度创造价值。根据该模型中的关联分析,根据组合学原理,结合 MapReduce 思想,提出基于大数据的频繁项集挖掘算法(Frequent Itemset Mining Based on Big data, FIMBB)。该算法只扫描一次原始事务数据库来完成整个频繁项集的挖掘过程;

利用大数据中的 MapReduce 平台并行挖掘出最终的所有目标频繁项集,整个流程采用了分布式和并行的思想,挖掘效率得到有效提高。

## 1 基于大数据的智能调度多源异构数据分析模型

智能调度大数据分析的主要思想为使用适当的大数据工具,抽取和集成多源异构数据,按照分析需求的统一格式存储预处理后的数据,采用数据分析和数据挖掘技术对存储的数据进行分析和深度挖掘,以提取出隐藏在数据中的知识,并根据智能调度的新需求,形成新的智能应用。

### 1.1 智能调度多源异构数据特点分析

调度环节的数据在传统电网基础上,数据来源、种类、规模都有了极大的扩充和丰富<sup>[10]</sup>,这些来自于不同系统的数据彼此之间有一定的关联性,不完全独立,这些数据结构复杂、数据量很大,彼此之间存在着复杂的关系。根据大数据的基本特征和电网调度的具体特点,智能调度多源异构数据具有以下特征:

(1)数据来自各调度中心,每个调度中心的数据又来源于多个系统,包括 SCADA、EMS、WAMS、AMI、OMS、GIS 等。每个系统采集到的数据模型、格式、特点不完全相同。

(2)数据规模大,维度多,实时性强。晋城供电公司 SCADA 系统大概总共有 80 000 个遥测点,采样间隔按 4 s 计算,每年将产生 11.014 TB 的数据。具体计算公式为:

$$11.014 \text{ TB} = (12 \text{ 字节/帧} \times 0.4 \text{ 帧/s} \times 80 \text{ 000 遥测点} \times 86 \text{ 400 s/天} \times 365 \text{ 天})/2^{40}$$

(3)数据的真实性和安全性高<sup>[11]</sup>。高质量电网调度数据对于数据分析和挖掘至关重要;调度是电网的中枢神经,数据的安全性是电网稳定、安全和可靠运行的前提条件。

(4)数据源之间的关联性强,集成全面分析产生的结果具有很大的经济和社会价值。例如,负荷预测<sup>[12]</sup>是智能调度中的一个关键应用,其预测主要以负荷数据为主,但是负荷预测与气象、地理、人口、经济等方面的数据有一定关联,若利用大数据技术,将这些相关的数据源进行全面负荷预测,将为电力用户创造极大的价值。

### 1.2 基于大数据的多源异构数据分析模型建立

与传统数据分析的主要区别在于:智能调度中大数据分析的数据往往包括大量的结构化、半结构化和非结构化数据。从数据来源到数据应用整个数据分析过程中,每个环节均能利用大数据处理平台 Hadoop、MapReduce<sup>[12]</sup>等方式进行并行处理。根据智能调度多

源异构数据的特征和大数据思想,建立了基于大数据的智能调度多源异构数据分析模型,如图 1 所示。

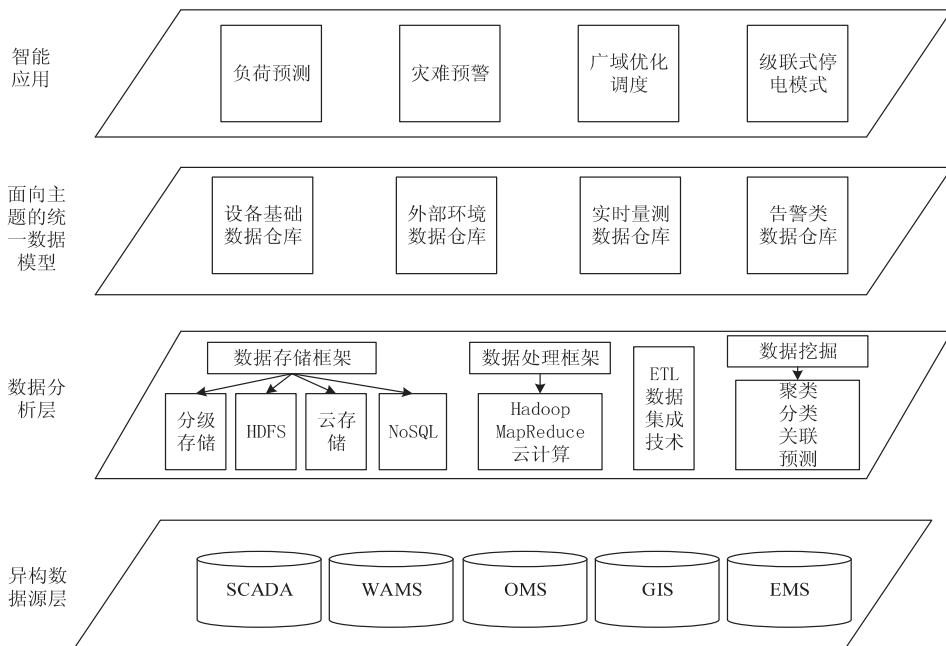


图 1 基于智能调度多源异构数据分析模型

图 1 描述了将电网智能调度的数据转化为对调度管理和决策有益的知识的过 程,打破了调度控制系统中各系统之间的信息孤岛。在异构数据源层,SCADA、WAMS 等系统都会产生海量数据,这些系统彼此不相同,数据类型复杂,因此需要首先对数据源的数据进行 ETL 预处理,保证数据质量及可靠性。NoSQL<sup>[13]</sup> 数据库技术是一种分布式数据存储方式,具有良好的可扩展性,解决了海量数据的存储难题。其中代表性的包括 Google 的 BigTable 和 Amazon 的 Dynamo 等。在云计算<sup>[14]</sup> 平台上,完成智能调度海量信息的可靠存储和快速并行处理。对存储后的数据通过数据挖掘等数据分析技术,将广泛的异构数据分类,这样多源异构数据源通过数据预处理和分析挖掘转化成了面向主题的、集成的调度全景大数据,如设备基础数据仓库、告警类数据仓库等统一数据模型,从而为系统提供全面的数据共享。将各类电网内外部数据和相应的调度业务数据进行结合,形成新的智能调度大数据应用场景。关联分析在智能调度数据分析挖掘中具有广泛应用,分析历史故障数据,找出故障之间的相关性,为快速找出根源故障,提供故障预测参考。

## 2 FIMBB 算法

设  $I = \{i_1, i_2, \dots, i_n\}$  为整个事务数据库中的全部项的集合,  $T = \{T_1, T_2, \dots, T_m\}$  为原始事务数据库中的  $m$  条记录,每条记录包含  $I$  中的若干项,  $T_k = \{t_1, t_2, \dots, t_q\}$  为其中的一条事务记录,  $\forall t_i \in T_k$ , 必有  $\forall t_i \in I$ 。设  $T'_k$  为事务记录  $T_k$  中的所有项组合构成的集合,  $I'$  为事务数据库中所有项组合构成的集合,则  $T'_k \subset I'$ 。

假设  $\emptyset \subset T'_k$ , 则  $T'_k$  中的元素总数为  $2^q$ ,  $q$  为  $T_k$  中项的个数。 $T'_k$  中的元素是频繁项集挖掘过程中可能出现的项集组合的可能性,频繁  $k$  项集指满足用户定义的最小支持度的  $k$  项集。例如设  $I = \{2, 5, 6, 8\}$ , 则  $I$  中总共有 4 项,组合结果为:  $I' = \{2, 5, 6, 8, 25, 26, 28, 56, 58, 68, 256, 258, 268, 568, 2\ 568\}$ 。则  $T_k = \{2, 5, 6\}$ ,  $T_k$  中的项通过组合得到集合  $T'_k = \{2, 5, 6, 25, 26, 56, 256\}$ , 则  $T'_k \subseteq I'$ 。对于海量的电网调度中故障事务数据库,MapReduce 利用多个 Map 和 Reduce 函数对事务数据进行频繁项集挖掘,以提高频繁项集挖掘的效率。FIMBB 算法的原理如图 2 所示。

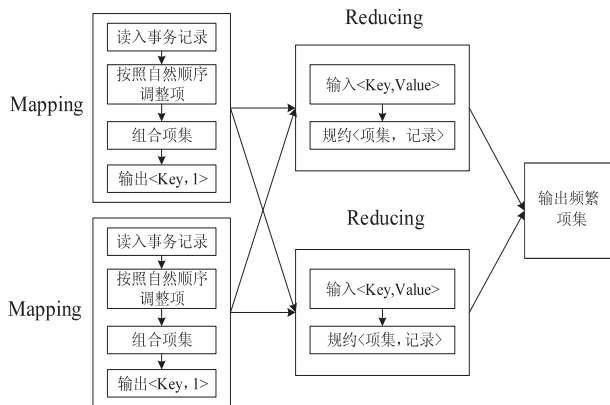


图 2 FIMBB 算法流程

Mapping 阶段算法的主要思想是:首先将输入的故障事务数据转化成程序中的统一格式,设置频繁项集挖掘的最小支持度阈值和  $k$  频繁项集的上限;读取每一条处理后的事务数据,将每条数据在满足项数上限的条件下组合事务,输出  $\langle \text{Key}, \text{Value} \rangle$ 。

算法主要步骤为:



- (1)将电网故障事务数据库的事务记录逐条读入;
  - (2)每条事务记录按照自然项进行处理;
  - (3)根据组合学原理,完成调整后的事务中项的组合;
  - (4)根据项的组合集合以〈Key,Value〉的形式输出,其中Key为事务记录项的组合,Value为1。
- 在 Reducing 阶段,将 Map 阶段的输出当作输入,合并相同的项集的计数。具体步骤为:当读到非空项集时,将项集的计数累加,然后统计其支持度;如果其支持度大于等于其最小支持度阈值,输出该项集。

3 实验与结果分析

实验分析中,采用人工随机电网调度中的故障信息事务集 Datafile 进行实验,其中故障事务中包含 10 个不同的故障信息项,并和 Apriori、FIMM 算法进行性能比较。FIMBB 算法是基于台式机搭建的 Hadoop 平台,该平台由三台计算机集群组成。其中,两台机器作为 DataNodes 和 TaskTrackers,这两台计算机配置了 N3700 核心处理器(主频 1.6 GHz)和 4 GB 内存;第三台计算机作为 NameNode 和 JobTracker,其配置了 G3260 双核处理器(主频 3.3 GHz)和 4 GB 内存。网络环境为同一局域网。实验结果如图 3 和图 4 所示。

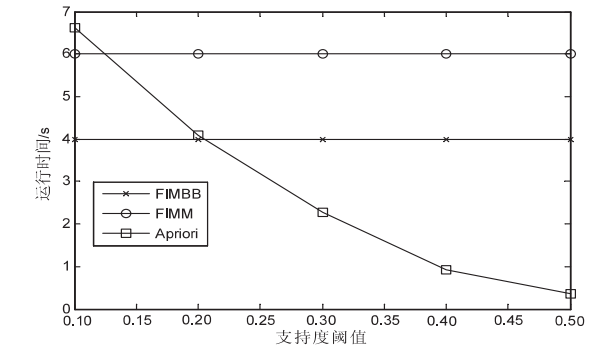


图 3 三种算法运行时间随着支持度阈值的变化趋势

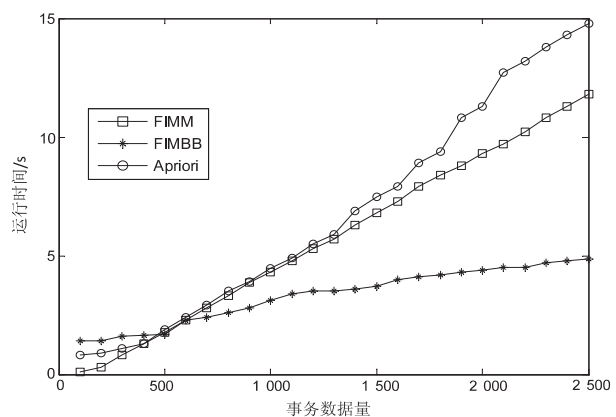


图 4 运行时间随事务数据量的变化趋势

从图 3 可以看到,FIMM 和 FIMBB 算法的运行时间基本不受支持度阈值的影响,而 Apriori 算法的运行

时间和支持度阈值的变化有很大关系,原因在于随着支持度的增大,Apriori 算法产生的候选项集数目变少,从候选集中找出真正的频繁项集相应所需的时间就变少。

从图 4 可以看到,FIMM 算法运行时间与事务数据量大小基本呈线性关系:当事务数据量小于 600 时,FIMM 的运行时间小于 FIMBB;当事务数大于 600 时,FIMBB 运行效率明显高于 FIMM。因为 FIMBB 算法在事务数据量较少时花费在配置运行环境和节点间通信上的时间占很大比例,当事务数据量较大时,FIMBB 算法的运行效率具有很大优势。在事务数较多时,Apriori 算法的运行效率明显要低于其他两种算法,原因在于事务数据量越大,Apriori 算法将产生大量的候选集且需要多次扫描原始事务数据库,因此耗时较多。当对不同数据量的事务进行挖掘时,FIMBB 算法的加速比如表 1 所示。

表 1 FIMBB 加速比

机器台数	数据量			
	1 000	10 000	100 000	200 000
2	1.14	1.27	1.33	1.35
3	1.12	1.71	1.76	1.82

根据理想加速比公式得到集群中机器台数分别为 2 和 3 的理想加速比为 2 和 3。从表 1 可以看出,在实际运行中得到的加速比往往和理想的差别很大。主要是由每台机器的硬件性能不完全相同造成的。另外,加速比计算公式为:

$$R = \frac{T_q + c_1}{T_s + c_2}$$

其中, $T_q$ 、 $T_s$  分别为集群和单机运行算法的时间开销; $c_1$ 、 $c_2$  为两者的系统开销。当事务数据量很大时,计算加速比可忽略系统开销。

从以上分析可知:对于大量的事务数据,FIMBB 算法的性能优于 FIMM 和 Apriori 算法,且具有很好的可扩展性,适用于智能调度中大量故障事务的频繁项集挖掘。

4 结束语

调度数据分析和处理是实现智能调度的关键,在分析智能调度数据特点的基础上,根据智能调度大数据的需求,构建基于大数据的智能调度多源异构数据分析模型,实现了通过大数据挖掘技术,将调度控制系统中的多源异构数据转化成智能调度的有价值信息。FIMBB 算法是一种针对大量电网调度事务数据的频繁项集挖掘算法。该算法将分布式计算的思想引入挖掘频繁项集中。

根据组合学原理,利用 MapReduce 扫描一次数据库从原始事务数据库中完成频繁项集的整体挖掘过程;且在支持度阈值改变的情况下无需重新扫描数据库进行挖掘,提高了频繁项集的挖掘效率。实验结果表明,该算法不受支持度阈值的影响,且对于大量事务数据,运行效率高,适用于智能调度大数据的关联分析。大数据在智能调度中的应用价值不可估量,但是,要加速智能调度化的进程,需要在多源数据融合和全景数据深度分析方面有所突破。

#### 参考文献:

- [1] 刘振亚. 智能电网技术[M]. 北京:中国电力出版社,2010.
- [2] 辛耀中,石俊杰,周京阳,等. 智能电网调度控制系统现状与技术展望[J]. 电力系统自动化,2015,39(1):2-8.
- [3] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases[C]//Proceedings of the ACM SIGMOD conference on management of data. Washington, D C:ACM,1993:207-216.
- [4] Han Jiawei,Pei Jian,Yin Yiwen. Mining frequent patterns without candidate generation [C]//Proceedings of the ACM SIGMOD conference on management of data. Dallas, TX:ACM,2000:1-12.
- [5] Baralis E, Cerquitelli T, Chiusano S, et al. Scalable out-of-core itemset mining[J]. Information Sciences,2015,293(4):146-162.
- [6] Baralis E, Cerquitelli T, Chiusano S. A persistent HY-Tree to efficiently support itemset mining on large datasets[C]//Proceedings of the 2010 ACM symposium on applied computing. New York:ACM,2010:1060-1064.
- [7] Adnan M, Alhajj R. DRFP-tree: disk-resident frequent pattern tree[J]. Applied Intelligence,2009,30(2):84-97.
- [8] Buehrer G, Parthasarathy S, Ghoting A. Out-of-core frequent pattern mining on a commodity PC [C]//Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. New York:ACM,2006:86-95.
- [9] 张东霞,苗新,刘丽平,等. 智能电网大数据技术发展研究[J]. 中国电机工程学报,2015,35(1):2-12.
- [10] 宋亚奇,周国亮,朱永利. 智能电网大数据处理技术现状与挑战[J]. 电网技术,2013,37(4):927-935.
- [11] 彭小圣,邓迪元,程时杰,等. 面向智能电网应用的电力大数据关键技术[J]. 中国电机工程学报,2015,35(3):503-511.
- [12] 李建江,崔健,王聃,等. MapReduce 并行编程模型研究综述[J]. 电子学报,2011,39(11):2635-2642.
- [13] 孟小峰,慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展,2013,50(1):146-169.
- [14] 吴凯峰,刘万涛,李彦虎,等. 基于云计算的电力大数据分析技术与应用[J]. 中国电力,2015(2):111-116.
- [15] Trentin E, Gori M. Robust combination of neural networks and hidden Markov models for speech recognition[J]. IEEE Transactions on Neural Networks,2003,14(6):1519-1531.
- [16] Hong K K, Rose R C. Cepstrum-domain model combination based on decomposition of speech and noise for noisy speech recognition[C]//IEEE international conference on acoustics, speech, and signal processing. [s. l.]:IEEE,2002:209-212.
- [17] Songhita M, Tusharkanti D, Partha S, et al. Comparison of MFCC and LPCC for a fixed phrase speaker verification system, time complexity and failure analysis[C]//International conference on circuit, power and computing technologies. [s. l.]:[s. n.],2015:1-4.
- [18] Yuan Y J, Zhao P H, Zhou Q. Research of speaker recognition based on combination of LPCC and MFCC[C]//International conference on intelligent computing and intelligent system. [s. l.]:[s. n.],2010:765-767.
- [19] Zhu J C, Liu Z L. Analysis of hybrid feature research based on extraction LPCC and MFCC [C]//10th international conference on computational intelligence and security. [s. l.]:[s. n.],2014:732-735.
- [20] Kopparapu S K, Laxminarayana M. Choice of Mel filter bank in computing MFCC of a resampled speech[C]//10th international conference on information sciences signal processing and their applications. [s. l.]:[s. n.],2010:121-124.
- [21] 周萍,李晓盼,李杰,等. 混合 MFCC 特征参数应用于语音情感识别[J]. 计算机测量与控制,2013,21(7):1966-1968.
- [22] 庞程,李晓飞,刘宏. 基于 MFCC 与基频特征贡献度识别说话人性别[J]. 华中科技大学学报:自然科学版,2013(S1):108-111.
- [23] 沈燕,肖仲喆,李冰洁,等. 采用 GW-MFCC 模型空间参数的语音情感识别[J]. 计算机工程与应用,2015,51(10):219-222.
- [24] 张家騄. 论语音技术的发展[J]. 声学学报,2004,29(3):193-199.
- [25] Watanabe A. Formant estimation method using inverse-filter control[J]. IEEE Transactions on Audio Processing,2001,9(4):317-326.
- [26] Rao P, Barman A D. Speech formant frequency estimation: evaluating a nonstationary analysis method[J]. Signal Processing,2000,80(8):1655-1667.
- [27] 韩志艳,伦淑娟,王健. 基于遗传小波神经网络的语音情感识别[J]. 计算机技术与发展,2013,23(1):75-78.
- [28] 韩志艳,伦淑娟,王健. 语音信号鲁棒特征提取及可视化技术研究[M]. 沈阳:东北大学出版社,2012.