分类号_____　　密　级_____
U D C _____　　编　号_____

清　华　大　学

博　士　后　研　究　报　告

电能质量数据高级分析技术研究

许　延　祥

清　华　大　学 (北京)

2014 年 07 月

电能质量数据高级分析技术研究

RESEARCH ON ADVANCED ANALYSIS TECHNOLOGIES OF POWER QUALITY DATA

博　士　后　姓　名:　　许延祥

流动站（一级学科）名称:　　控制科学与工程

专　业（二级学科）名称:　　控制理论与控制工程

合　　作　　导　　师:　　曹军威（研究员）

研究工作起止日期　2012 年 07 月　一　2014 年 07 月

中文摘要

摘要：现代电力系统中，电力电子设备的应用越来越广泛,各种非线性、冲击性、波动性负载也大量增加，使电力系统所遭受的电能质量污染也日趋严重。同时信息科技的发展则对电能质量及供电可靠性提出了更高的要求。电能质量问题已引起电力系统及电力用户的广泛关注，加强电能质量技术监督已经成为供用电系统的共识。在此基础上，国内诸多电力企业、用户相继建设了电能质量监测系统， 对电能质量特征指标进行不间断的实时监测，其监测数据由数据库统一管理。

得到电能质量监测数据只是进行电能质量分析的第一步，如何通过监测数据判断被监测对象是否存在问题，存在哪些问题，才是关键所在。作为电能质量监测管理人员，面对海量监测数据如何分析是目前电能质量监测系统能否发挥其应有作用的客观技术难题。

对于电能质量数据的初级分析手段已经被广泛采用，但是为了进一步提高电能质量，优化电力系统，我们需要向更高级的数据分析手段及信息挖掘技术迈进。从分析目标来看，对于电能质量数据我们不再满足于统计已有情况，而希望向预测未来发展迈进；不再满足于对单个监测节点及线路的电能质量进行评估，而希望能够分析并发现某个电网局域中各节点的影响关系，甚至发现问题源头；我们不再满足于从数据的统计中知道一些我们预期看到的规律，而希望海量的数据能告诉我们更多的我们不曾想到的但却对我们有价值的信息。

本文面向上述电能质量高级分析问题，引入数据挖掘技术：1）针对电能质量监测采集的大数据，提出 分布式数据挖掘的技术框架；2）针对电力系统产生流数据的特点，提出使用流计算框架；3）针对电能质量问题的预测要求，归纳总结了统计回归与时间序列分析技术；4）针对复杂决策需要，归纳了分类技术，并介绍了9种典型的分类算法；5）针对发现未知分布的需要，归纳了聚类技术，并介绍了5类典型的聚类算法； 6）针对在局域中快速定位问题的需要，提出并联规则分析方法；7）针对高级数据分析过程中人机交互需要，归纳总结了数据可视化技术。

在面向区域电网进行节点间影响关系分析方面，本研究提出两种创新性的方法：1）将关联规则分析法应用于多相邻节点在某方面电能质量问题上的相互影响关系分析。其实现的挑战性问题是在现有数据基础上，生成多节点在被观察问题上的时间轴对齐的数据，以便应用现有数据挖掘算法。本文对来自实际系统的数据进行了实验分

析，采用的关联规则算法为FP-growth算法。2)提出一种利用全网电能质量监测数据，通过计算各监测节点之间出现谐波的顺序关系来定位谐波源的方法。在当前电能质量监测仪广泛部署应用的情况下，该方法利用监测数据记录时间连贯的特性，以迭代扫描方式从中提取各节点的谐波序列，再计算两两节点间包含5种预定义的序列关系的比例，进而定位出疑似谐波源位置。以生产系统中数据实例为验证，本文方法定位的谐波源与生产环境实际所知谐波源位置一致。因而，该方法在不进行额外仪器设备投入的前提下，利用历史数据，能够实现快速有效定位电网中谐波源的目的。

总之，充分利用数据挖掘技术有望从大规模、高维的电能质量监测数据中提取出隐藏的模式和规则，为电力系统决策者提供决策支持具有重要的研究价值。本报告正是为了全面研究数据挖掘技术应用于电能质量高级分析的各种技术可能性。

关键词：电力系统分析；电能质量；数据挖掘； 谐波源定位；电压事件分析

ABSTRACT

ABSTRACT: With more and more power electronic devices widely deployed in modern power systems, various non-linear impactive undulant loads increased load. Power systems are increasingly being suffered power quality pollution. At the same time the development of information technology expect high-level power quality and more reliability of power supply. Power quality issues have caused widespread concern in the power systems and power users. It has become the consensus for electricity systems and uers to enhance power quality technical supervision . On this basis, many domestic power companies and users have built power quality monitoring systems to continuously real-time monitor power quality indicators, and the monitor data is centrally managed by a database.

It's just the first step of the analysis of power quality to obtain power quality monitoring data.The key issue is to judge what problem exists on the monitored object. It's an objective difficulty for power quality monitoring and management staff to analysis massive data analysis utilizing the current power monitoring system.

Many primary power quality questions have been widely handled, but in order to further improve the power quality, we need more advanced data analysis methods and data mining technology. We need predicting the future development of power quality than only knowing the existing situation in the statistics. We hope to be able to analyze the affect relationship between nodes in a local power grid and even to finding disturbance source than only being able to analyze monitoring node one by one.We are no longer satisfied to know some of the laws that we expect to see from the statistics, and we want the vast amounts of data to tell us more valuable information that we haven’t being realized.

This article introduces data mining technology for the above advanced power quality analysis problems. 1)For the massive data, distributed data mining technique framework is proposed; 2) For the streaming data in power system, stream computing framework is proposed; 3) For the requirements of power quality problems he prediction, the statistical regression and time series analysis are introduced; 4) For the need of complex decision-making, classification techniques are summarized and nine kinds of typical classification algorithms are presented; 5) For discovery of unknown distribution, the clustering techniques are summed up and five types of typical clustering algorithm are described; 6) For the need to quickly locate power quality problems in a grid area, the

association rule analysis method is applied in this problem; 7) For the requires on human-computer interaction in advanced data analysis, data visualization techniques are summarized.

In the analysis of the relationship between nodes in a regional power grid, this study proposes two innovative ways: 1) With the current widely deployed power quality monitors, this paper presents a novel harmonic localization method of computing the harmonic sequential relationships between the monitored nodes utilizing power quality data from the whole grid. Utilizing the time-continuity of the monitoring data, the method can extract the harmonic sequences of each node by iteratively scanning the record set. Then the method can calculate the proportions of 5 kinds of predefined sequence relationships between any two nodes, and then locate the suspected harmonic sources based on the proportion. Taken the data from real system for verification, the harmonic source calculated out by the method is consistent with the known one in actual system. Therefore, without extra investment of instrument and equipment, using historical data, the method can realize the rapid and effective positioning of harmonic source in grid. 2) This paper presents a method to analyze the mutual influence of the voltage events generated from the different nodes and then locate the voltage disturbance sources in a grid based on massive existing power quality monitoring data and association rule algorithm. This paper focuses on the implementation of transforming a batch of actual power quality data from a single-node sequential list into a multi-node parallel two-dimensional table. After that, we chose the PF-growth algorithm with some appropriate parameters to calculate the affecting relationship on voltage events between the nodes. Relative to the traditional methods, such as system simulation and matrix calculations, this method has low cost, fast and efficient computing features.

Full application of data mining technologies in power quality analysis is expected to extract hidden patterns and rules from large-scale, high-dimensional power quality monitoring data, and provide decision support for decision-makers in power system. This report is a comprehensive research on the technical possibilities of applying data mining technologies for power quality advanced analysis.

KEYWORDS： Power grid analysis; Pwer Quality; Data mining; Harmonic source localization; Voltage disturbance source

目录

图目录

表目录

前言

现代电力系统中，电力电子设备的应用越来越广泛,各种非线性、冲击性、波动性负载也大量增加，使电力系统所遭受的电能质量污染也日趋严重。同时信息科技的发展则对电能质量及供电可靠性提出了更高的要求。电能质量问题已引起电力系统及电力用户的广泛关注[1]，加强电能质量技术监督已经成为供用电系统的共识。在此基础上，国内诸多电力企业、用户相继建设了电能质量监测系统，对电能质量特征指标进行不间断的实时监测，其监测数据由数据库统一管理。

得到电能质量监测数据只是进行电能质量分析的第一步，如何通过监测数据判断被监测对象是否存在问题，存在哪些问题，才是关键所在。作为电能质量监测管理人员，面对海量监测数据如何分析是目前电能质量监测系统能否发挥其应有作用的客观技术难题。

监测仪器除了要根据需求记录电压和电流实时波形数据、电能质量指标和超标数据外，还要记录日、月、年趋势指标数据等分析统计指标；其次，每种电能

质量问题所涉及的特征量都非常多，整个监测系统中，监测点往往不只一个，所记录的数据包括实时的和历史的，因此，电能质量监测数据已成为了包含多监测节点涵盖多方面问题的海量数据[2]。大量的数据对进行电能质量事件预测、故障辨识、干扰源识别和实时控制形成了巨大的挑战。在这种情况下，迫切需要数据分析的智能方法，能从电能质量扰动数据中发现有用的信息。

对于电能质量数据的初级分析手段已经被广泛采用，但是为了进一步提高电能质量，优化电力系统，我们需要向更高级的数据分析手段及信息挖掘技术迈进。从分析目标来看，对于电能质量数据我们不再满足于统计已有情况，而希望向预测未来发展迈进；不再满足于对单个监测节点及线路的电能质量进行评估，而希望能够分析并发现某个电网局域中各节点的影响关系，甚至发现问题源头；我们不再满足于从数据的统计中知道一些我们预期看到的规律，而希望海量的数据能告诉我们更多的我们不曾想到的但却对我们有价值的信息。

数据挖掘技术在这一需要下被考虑引入进来。数据挖掘(Data Mining)技术，也称为数据库知识发现(KDD，Knowledge Discovery in Database)。随着信息社会中计算机技术、存储技术及互联网的快速发展，各行业数据库的规模日益增大。传统数据库系统技术例如录入、查询、统计等操作型处理，获得的信息仅仅是整个数据库所包含信息的一部分，无法发现数据中存在的隐藏关系、规则以及未来的发展趋 势。在此背景下，传统的数据分析及统计学理论，结合现代的数据库技术和人工智能算法，产生了数据挖掘这门跨学科技术，用以发现隐含在大量数据中、先前未知的、对决策有潜在价值的规律和知识。利用数据挖掘技术，我们能够对电力系统内采集的大量的电能质量监测数据进行去芜存菁，发现有用数据，降低数据处理量。在此基础上，从电能质量监测系统中提取出用户电压波形，可以利用数据挖掘技术的分类功能自动识别电能扰动事件并进行分类，为电能质量治理提供依据；基于电能质量问题数据可以建立系统安全评估决策树，快速判断系统运行状态并提出相应的控制策略；在区域监测数据基础上可以进行电能质量区域评估；结合网络拓扑信息可以进行谐波源识别；基于电能质量问题的历史数据可以进行问题预测及预警。

电力系统是现代社会中典型的“巨系统”，其以多节点迸发的方式持续、高速地产生着海量的电能质量数据。传统的数据挖掘方法多是对批量截取出来的数据进行着基于单机内存计算的处理。面向电能质量数据，我们有必要引入并行数

据挖掘计算方法、流数据处理等新技术。

总之，充分利用数据挖掘技术有望从大规模、高维的电能质量监测数据中提取出隐藏的模式和规则，为电力系统决策者提供决策支持具有重要的研究价值。本报告正是为了全面研究数据挖掘技术应用于电能质量高级分析的各种技术可能性，给出可能应用于电能质量数据高级分析的技术汇总与梳理，也提出了面向区域电网进行整体分析的两种新方法。

1 绪论

1.1 研究背景

本研究是江苏省电力公司所承担的国家电网公司科学技术项目“电能质量高级分析关键技术研究及应用”的子课题“电能质量数据集成与信息挖掘技术”的成果之一。

电能质量高级分析关键技术研究及应用的研究目标如图 1 所示。

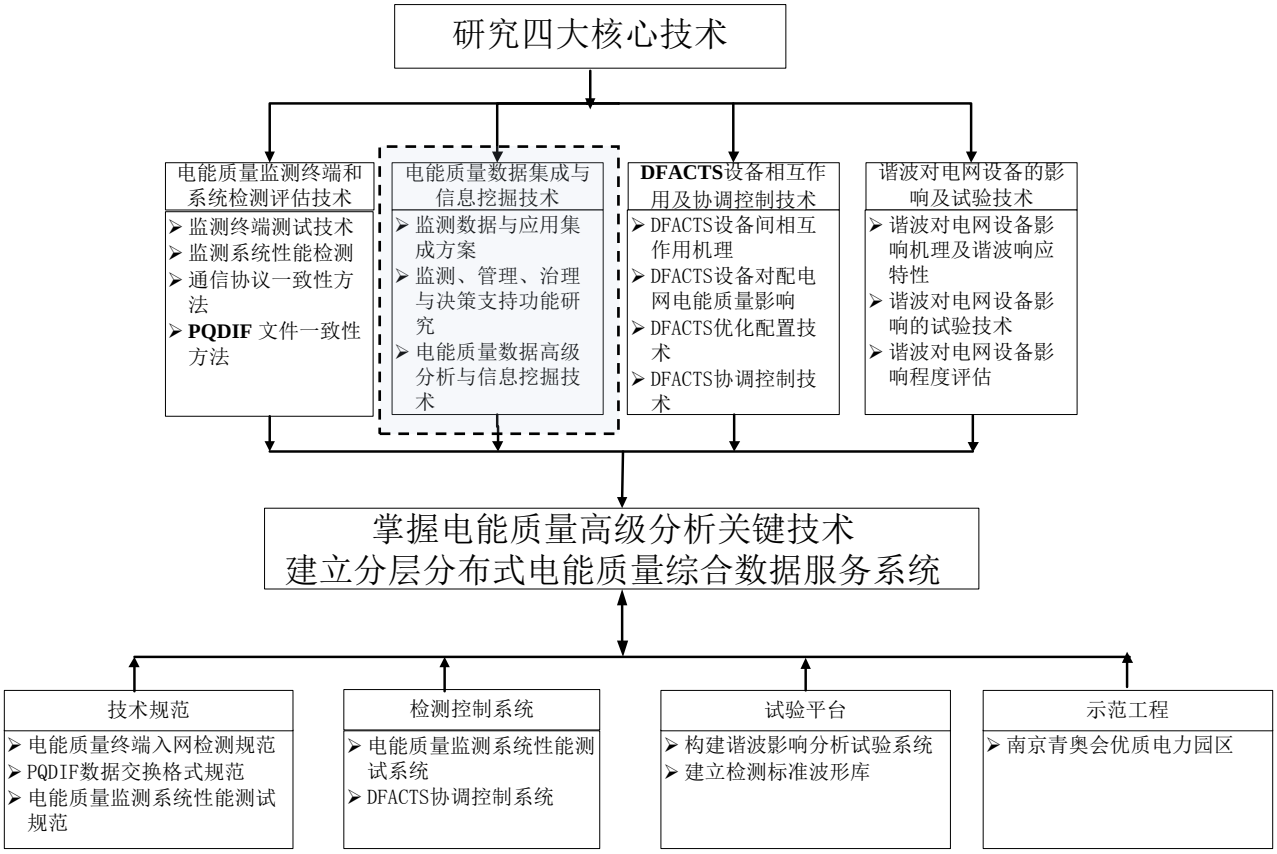


图 1 电能质量高级分析关键技术研究及应用研究目标

其中，电能质量数据挖掘和高级应用分析的目的是采用各种先进的数据挖掘和数据分析技术，结合智能算法实现电能质量的分析和处理，实现电能质量综合评估、预警、扰动源识别等高级应用，同时为电能质量的治理提供了有效的保障。具体目标如下：

（1）基于数据挖掘的电能质量数据预处理

高质量的决策必须依赖高质量的数据，对电能质量采集数据进行预处理至关重要。预处理一般包括数据压缩和数据去噪两部分。通过分箱、聚类、回归等方

法对数据进行清洗和降噪，预处理后的数据不仅为电能质量监测、治理提供可用可行的第一手资料，且能够为今后电网规划、经济区、工业规划等提供参考资料。

（2）电能质量综合评估

研究基于灰色理论和组合线性加权的电能质量综合评估算法，对电能质量监测数据进行分级分区的单指标以及综合评估，掌握电网侧和用户侧的全网各级各区电能质量状况，及时发现电能质量较差区域，为电能质量治理提供有效信息。

（3）电能质量预警

通过分析各个站点的电能质量监测数据，及时发现各种电能质量的异常变化，提前发现电网电能质量出现的隐患，避免故障的扩大；同时，向故障区域内易受电能质量异常变化影响的用户发出预警信息，与用户共享电能质量监测数据，使用户做出及时的用电决定，减小电能质量异常变化的不利影响。

（4）电能质量扰动源识别

电能质量扰动源识别可以明确供用电双方的责任，判别扰动是由电力系统引起还是由用户引起。找到引起的扰动源，以便采用有效的措施来消除和减轻扰动的影响。由于监测到的电能质量相关数据十分庞大，可以使用关联分析的方法挖掘出电能质量问题相关扰动电流、扰动电压与扰动源之间的关系，并对各种电能质量扰动提取特征向量；通过敏感因子确定监测点的数量、分布及强度,利用基于最小二乘支持向量机的扰动源定位并准确计算出扰动源注入的扰动电流大小。与人工智能技术相结合，可以实现扰动类型的自动识别，将有助于分析扰动产生的原因，找出合理的应对措施。

本子课题研究成果最终形成 4 个报告，3 篇论文和 1 个系统。本报告即是 4 个报告之一。

1.2 问题概述及意义

现代电力系统中，电力电子设备的应用越来越广泛,各种非线性、冲击性、波动性负载也大量增加，使电力系统所遭受的电能质量污染也日趋严重。同时信息科技的发展则对电能质量及供电可靠性提出了更高的要求。为了提高电能质量，首先，电能质量定义中所涉及到的影响电力用户设备或系统不能正常工作的电压、频率或电流偏离标称值的情况都应当被监测和记录。监测仪器除了要根据需求记录电压和电流实时波形数据、电能质量指标和超标数据外，还要记录日、月、年

趋势指标数据等分析统计指标；其次，每种电能质量问题所涉及的特征量都非常多，整个监测系统中，监测点往往不只一个，所记录的数据包括实时的和历史的，因此，电能质量监测数据无疑已成为了海量数据[3] 。

大量的数据对进行电能质量事件预测、故障辨识、干扰源识别和实时控制形成了巨大的挑战。在这种情况下，迫切需要数据分析的智能方法，能从电能质量扰动数据中发现有用的信息。数据挖掘技术正是在这一需要下被考虑引入进来。

数据挖掘(Data Mining)技术，也称为数据库知识发现(KDD，Knowledge Discovery in Database)。随着信息社会中计算机技术、存储技术及互联网的快速发展，各行业数据库的规模日益增大。传统数据库系统技术例如录入、查询、统计等操作型处理，获得的信息仅仅是整个数据库所包含信息的一部分，无法发现数据中存在的隐藏关系、规则以及未来的发展趋 势。在此背景下，传统的数据分析及统计学理论，结合现代的数据库技术和人工智能算法，产生了数据挖掘这门跨学科技术，用以发现隐含在大量数据中、先前未知的、对决策有潜在价值的规律和知识。

利用数据挖掘技术，我们能够对电力系统内采集的大量的电能质量监测数据进行去芜存菁，发现有用数据，降低数据处理量。在此基础上，从电能质量监测系统中提取出用户电压波形，可以利用数据挖掘技术的分类功能自动识别电能扰动事件并进行分类，为电能质量治理提供依据；基于电能质量问题数据可以建立系统安全评估决策树，快速判断系统运行状态并提出相应的控制策略；在区域监测数据基础上可以进行电能质量区域评估；结合网络拓扑信息可以进行谐波源识别；基于电能质量问题的历史数据可以进行问题预测及预警。

总之，充分利用数据挖掘技术有望从大规模、高维的电能质量监测数据中提取出隐藏的模式和规则，为电力系统决策者提供决策支持具有重要的研究价值。本报告正是为了全面研究数据挖掘技术应用于电能质量分析的各种技术可能性。

1.3 研究路线

首先，本文对电能质量数据的内容和形式进行介绍，以此作为进一步分析的基础。

然后，由于本报告的研究主旨为面向电能质量数据的“高级分析”，本文从 5 个方向上阐述了高级分析相对初级分析的高级所在，即所分析问题的复杂性。这

五个方向是：从面向统计为主转向面向预测为主；从面向单个节点转向面向局域网络；从面向已知问题转向面向未知规律；从面向集中数据转向面向分布的数据；从面向批量数据的事后分析转向面向流数据的实时处理。在这五个方向的介绍中会以问题为主线，介绍所需的统计预测算法及数据挖掘算法。

最后，面向电力系统本身的复杂性，我们提出所采用的信息挖掘技术要能够并行处理海量的、流式的数据，因而在技术上我们引入介绍并行化数据挖掘框架、流数据计算框架，在此基础上，我们设计了基于并行挖掘分析的流数据处理框架。

具体路线如下图如示。

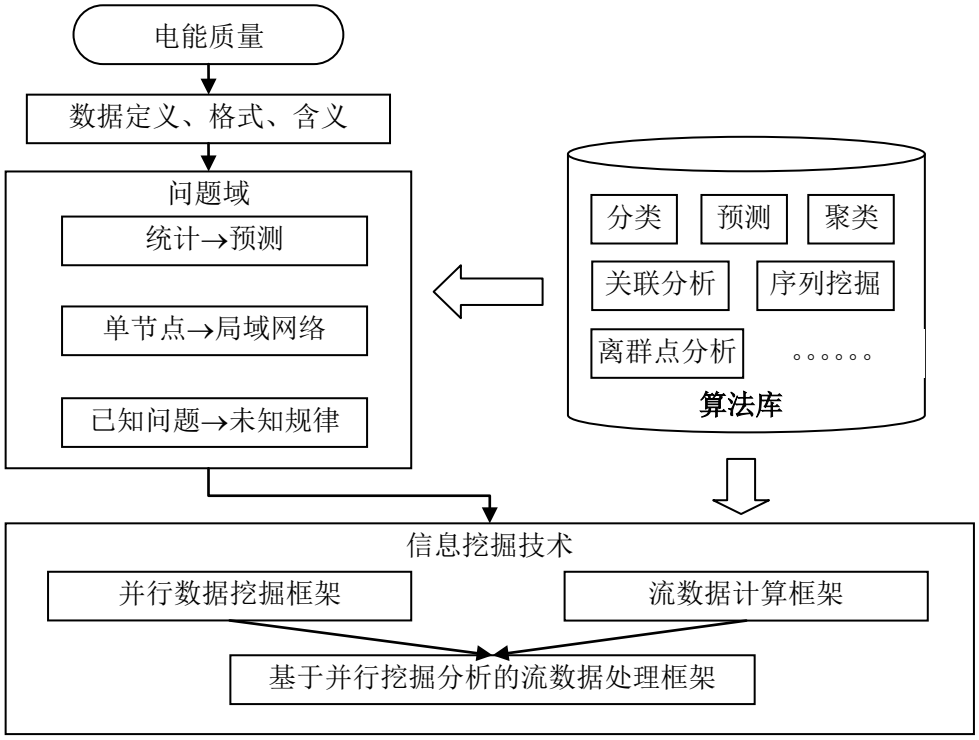


图 2 研究路线

本报告的章节组织也大体依照上述电能质量数据高级分析和信息挖掘技术的研究路线。

2 电能质量问题及数据

本章将明确电能质量的定义，并对电能质量监测办法和数据格式进行基本说明分析。

2.1 电能质量问题定义

电能质量是在电力系统的大环境下所定义产生的，因此首先需要从大电网的角度来看电能质量问题。

电力系统由发电、输电和配电/负荷 3 个环节组成。3 个子系统必须作为一个整体来一起运行。任何时刻，电力系统的发电功率必须和负荷功率相匹配。无论何时何地，对于系统中的负荷变化和可能发生的扰动，所有发电机必须保持同步运行（用 HVDC 等连接方式可以使互联系统非同步运行）。了解每个子系统很重要，了解它们的关联特性也很重要。其中电能质量以及由质量所反映的系统运行特性是重要的研究课题之一。

不间断地向用户供应质量（电压和频率）合乎规定的电能。对其性能要求通常以可靠性来表征。电力系统可靠性是按可接受的质量（Quality of power ）标准和所需电量（Quantity of energy）不间断地向电力用户提供电力和电量的能力的量度。Q×Q（电量，质量）是系统运行可靠性（充裕性和安全性）的基本表现。

电力系统有 6 种运行状态：

1）正常状态：能够保持系统充裕性和系统安全性的运行状态；

2）紧戒状态：可以通过调整控制恢复到正常状态；

3）紧急状态：某些母线电压或系统频率超越允许范围。允许损失部分负荷

4）极端紧急状态或严重事故状态：不至于造成大面积停电；

5）崩溃：系统稳定破坏，故障连锁反应，电压频率崩溃，大面积减负荷，系统解列，孤岛状态；

6）恢复过程：黑启动，“灰启动”。

在正常运行状态下，除了系统安全性之外，经济性也是一个重要的方面。在紧急状态下，电力系统的经济运行处于次要地位。在系统设计和运行中需要权衡和兼顾经济性（充裕性与经济性紧密关联，是一个重要方面）和安全性。

在时间上占主导地位的正常运行状态是电能质量得以保证的前提条件。或者

说，电能质量是在正常运行条件下才有评估意义的；正常运行是电力系统的大概率事件，时刻关注电能质量是十分必要的。

电能质量是一个覆盖诸多学科的复杂领域，而科技的发展和社会文明的进步，又赋予了电能质量许多新的内容。作用在电力用户设施和装备上的电能质量问题已经引起电工界的极大关注，以提出特定条件下的单体设备或工厂为对象的解决方法较为普遍。建立友好型电磁环境是多方面共同努力的结果，而施加在公用电网的电能质量问题相比较还是新问题（典型的有畸变条件下的功率理论与电能计量方法；电力系统运行状态与电能质量的关系；新能源和分布式电源的接入；电力系统的经济性与电能质量的市场化操作等），其理论和技术的支撑工作仍处于刚刚开始阶段[4]。

从微观来看，电能质量就是指通过公用电网供给用户端的交流电能的质量。理想状态的公用电网应以**恒定的频率**、**标准正弦波**和**额定电压**对用户供电。同时，在三相交流系统中，各相电压和电流的**幅值大小应相等**、**相位对称且相差 120 度**。但由于系统中的发电机、变压器和线路等设备非线性或不对称、负荷性质多变，加之调控手段不完善及运行操作、外来干扰和各种故障等原因，这种理想状态并不存在。因此，产生了电网运行电力设备和供用电环节中的各种问题，也就产生了电能质量的概念。

电能质量（power quality）最先来源于 1968 年一篇关于美国海军电子设备一电源规范研究的论文中。一般来说，电能质量是指优质供电，供电部门向用户（系统向负荷）提供合格的电力。广义上来说，系统的可靠性、绝缘及接地体的选择、供电中断、三相电压不平衡、电力电子设备对电网的干扰等造成用户设备故障或错误动作的任何电力问题都属于电能质量问题。狭义上来说，电能质量问题主要集中在波形畸变上，表现为电压、电流和频率的偏差。从技术方面考虑，电能质量是供用电双方共同保证的，两者之间存在相互作用和影响。

IEEE 标准化协调委员会正式采用“power quality”这一术语，并给出了技术定义：“合格电能质量的概念是指，给敏感设备提供的电力和设置的接地系统是均适合于该设备正常工作的”。此外，这一研究领域的许多文献和报告还给出了一些未经公认的术语和补充定义，如：电压质量、电流质量、供电质量、用电质量等。

由于所处立场不同，关注电能质量的角度不同，人们对电能质量的定义还未

能达成完全的共识，但是对其主要技术指标都有较为一致的认识。本文所考虑的电能质量问题主要参照了 IEEE 对电能质量的划分，共包括 5 大类 13 个电能质量指标。

2.1.1 暂态事件

暂态事件包括脉冲暂态扰动主要有两种：脉冲暂态和振荡暂态。

（1）脉冲暂态

脉冲暂态(impulse)，指处于稳态的电压或电流突发的、单极性变化，大多由雷电引起。

脉冲暂态用上升时间和衰减时间描述，见 表格 1。

表格 1 脉冲暂态特征

类别	波形特征	典型持续时间
纳秒级	5ns 上升	<50ns
微妙级	1μs 上升	50ns～1ms
毫秒级	0.1ms 上升	>1ms

（2）振荡暂态

振荡暂态，指处于稳态的电压或电流突发的、双极性变化。与脉冲暂态的区别在于扰动波形的极性，脉冲扰动为单极性者，振荡扰动为双极性。根据其频谱成分，振荡暂态又分为高、中、低频三种类型，见表格 2。

表格 2 振荡暂态特征

类别	频谱成分	典型持续时间	电压幅值(p.u.)
低频	<5KHz	0.3~0.5ms	0~4
中频	5~500KHz	20μs	0~8
高频	0.5~5MHz	5μs	0~4

2.1.2 电压偏差

电压偏差（voltage deviation）是供电系统在正常运行方式下，某一节点的实际电压与系统的标准电压之差对系统标称电压的百分数。其数学表达式如下：

$$\delta U=\frac{U_{re}-U_N}{U_N}\times 100\% \qquad (2\text{-}1)$$

式中：δU—电压偏差； U_{re}—实际电压（kV）；U_N—系统标称电压（kV）。

《GB12325-90 电能质量-供电电压允许偏差》中规定：35kV 及以上供电电压正负偏差的绝对值之和不超过额定电压的 10%；10kV 及以下三相供电电压允许偏差为额定电压的±7%；220kV 单相供电电压允许偏差为额定电压的+7%，-10%。

确定允许电压偏差是一个综合的技术经济问题，允许的电压偏差小，有利于用电设备的安全、经济运行，但须为此在电网中增添更多的无功电源和调压设备，需要更多的投入。反过来，如果扩大用电设备对电压的适应范围，提高设备在这方面的性能往往也需增加设备的投资。供电系统在正常运行时，负荷时刻发生着变化，系统的运行方式也经常改变，系统中各节点的电压随之发生改变，会偏离电压标称值。电压的这种变化是缓慢的，其每秒电压变化率小于标称电压的 1%。供电电压允许偏差是电能质量的一项基本指标，合理确定该偏差对于电气设备的制造和运行，对于电力系统安全和经济都有重要意义。系统无功功率不平衡是引起系统电压偏差的根本原因，无功功率越严重，电压偏差越大。另外供配电网络结构的不合理也能导致电压偏差。供配电线路输送距离过长，输送容量过大，导致截面过小等因素都会加大线路的电压损失，从而产生电压偏差。

电压偏差过大所产生的危害有：① 对用电设备的危害：当电压偏离标称电压较大时，用电设备的运行性能恶化，不仅运行效率降低，还可能由于过电压或过电流而损坏。 例如，当电压高于标称电压 5%时，白炽灯的寿命会减少 30%，当电压高于标称电压 10%时，白炽灯的寿命会减小一半，从而使白炽灯的损坏数量大大增加，当电压低于标称电压的 5%时，白炽灯的光通量减少 18%。当电压低于标称电压 10%时，光通量减少 30%，从而使照度显著降低。电压过低或过高都会使电动机的温升增加，若电动机长时间处于较大的电压偏差下运行，就可能烧坏电动机绕组，使绕组绝缘老化而缩短电动机的寿命。由于许多家用电器内部都装有动力装置，也即是各种类型的电动机，电压偏差过大同样会影响它们的使用效率和寿命，严重影响人们的正常生活。② 对电网的危害: 输电线路的输送功率受功率稳定极限的限制，而线路的静态稳定功率极限近似与线路的电压平方成正比。系统运行电压偏低，输电线路的功率极限大幅度降低，可能产生系统频率不稳定现象，甚至导致电力系统频率崩溃，造成系统解列。如果电力系统缺乏无功电源，可能产生系统电压不稳定现象，导致电压崩溃。

2.1.3 短时电压变化

短时电压变化包括电压中断（Interruptions）、暂升（Swell）、暂降（Sag）。根据其持续时间之长短，这三种电压变化又分别分为即时（Instantaneous）、瞬时（Momentary）、暂时（Temporary）三种状态，其频谱特征见表格 6。

表格 3 短时电压变化特征

类别		典型时长	电压幅值（p.u.）
即时	暂升	0.5~30 周	1.1~1.8
	暂降	0.5~30 周	0.1~0.9
瞬时	中断	0.5 周~3s	<0.1
	暂升	30 周~3s	1.1~1.4
	暂降	30 周~3s	0.1~0.9
暂时	中断	3s~1min	<0.1
	暂升	3s~1min	1.1~1.2
	暂降	3s~1min	0.1~0.9

（1）电压暂降

电压暂降或下跌是指供电电压有效值在短时间内突然下降又回升恢复的现象。目前各种资料描绘电压短时间下降还有“电压凹陷”和“电压骤降”两种说法。实际上这这些术语都是一回事，只是标准的规定值有所不同而已。在电网中这种现象的持续时间大多为 0.5～1.5s。

国际电气与电子工程师协会（IEEE）把这一现象称为电压凹陷(voltage sag),其定义为:供电系统中某点的工频电压均方根值突然下降至额定值的 10%~90%,并在随后的 10ms~1min 的短暂持续期后恢复正常。

国际电工委员会（IEC）将这一现象称为电压骤降(voltage dip),其定义为:供电系统中某点的工频电压均方根值突然下降至额定值的 1%~90%,并在随后的 10ms~1min 的短暂持续期后恢复正常。

电压暂降一般是由电网、变电设施的故障或负荷突然出现大的变化(如大功率设备启动等)所引起的。在某些情况下会出现两次或更多次连续的跌落或中断。电压暂降会使用户的次品率增大或生产停顿。

（2）电压骤升

电压上升(swell)、电压暂升是指工频条件下电压均方根值上升到 1.1---1.8 倍额定电压之间、持续时间为 0.5 周波至 1 分钟的短时间电压变动现象；它们的起因都是系统故障。

（3）供电中断

当供电电压降低到 0.1 倍额定电压以下，且持续时间不超过 1 分钟，我们认为发生了电压中断（Interruptions），电压中断是一种短时间电压变动现象；造成电压中断的原因可能是系统故障、用电设备故障或控制失灵等。

2.1.4 长时电压变化

长时电压变化包括：欠电压、过电压、停电等。

（1）欠电压

欠电压(under-voltage) 是指工频下交流电压均方根值降低，小于额定值的 10%，并且持续时间大于 1 分钟的长时间电压变动现象；引起欠电压的事件正好与过电压相反，某一大容量负荷的投入或某一电容器组的断开（无功严重不足引起的欠电压）都可能引起欠电压。

欠电压的危害有：①电压降低 10%~15%，发电机输出功率减小 5%~10%，发电机有功功率减小；②电压低至 70%及以下时可能会发生电网奔溃，造成大面积停电；③如果电压降低 20%，电机转矩减小 30%，电流增大 20%~30%，电机温升升高 15 度左右；④接触器，继电器吸力减小，可能导致触头烧毁；⑤电压过低电机无法启动，或者转速变慢，堵转甚至烧毁电机。

（2）过电压

过电压(over-voltage)是指电力系统在工频下交流电压均方根值升高，超过额定值的 10%，并且持续时间大于 1 分钟的长时间电压变动现象。过电压的出现通常是负荷投切的结果，例如：切断某一大容量负荷或向电容器组增能（无功补偿过剩导致的过电压）。过电压种类过电压分外过电压和内过电压两大类。

外过电压又称雷电过电压、大气过电压。由大气中的雷云对地面放电而引起的。分直击雷过电压和感应雷过电压两种。雷电过电压的持续时间约为几十微秒，具有脉冲的特性，故常称为雷电冲击波。直击雷过电压是雷闪直接击中电工设备导电部分时所出现的过电压。雷闪击中带电的导体 ，如架空输电线路导线，称为直接雷击。雷闪击中正常情况下处于接地状态的导体，如输电线路铁塔，使其电位升高以后又对带电的导体放电称为反击。直击雷过电压幅值可达上百万伏，会破坏电工设施绝缘，引起短路接地故障。感应雷过电压是雷闪击中电工设备附

近地面，在放电过程中由于空间电磁场的急剧变化而使未直接遭受雷击的电工设备（包括二次设备、通信设备）上感应出的过电压。因此，架空输电线路需架设避雷线和接地装置等进行防护。通常用线路耐雷水平和雷击跳闸率表示输电线路的防雷能力。

内过电压是电力系统内部运行方式发生改变而引起的过电压。有暂态过电压、操作过电压和谐振过电压。

暂态过电压是由于断路器操作或发生短路故障，使电力系统经历过渡过程以后重新达到某种暂时稳定的情况下所出现的过电压，又称工频电压升高。常见的有：①空载长线电容效应（费兰梯效应）。在工频电源作用下，由于远距离空载线路电容效应的积累，使沿线电压分布不等，末端电压最高。②不对称短路接地。三相输电线路 a 相短路接地故障时 ，b、c 相上的电压会升高。③甩负荷过电压，输电线路因发生故障而被迫突然甩掉负荷时，由于电源电动势尚未及时自动调节而引起的过电压。

操作过电压是由于进行断路器操作或发生突然短路而引起的衰减较快持续时间较短的过电压，常见的有：①空载线路合闸和重合闸过电压。②切除空载线路过电压。③切断空载变压器过电压。④弧光接地过电压。

谐振过电压是电力系统中电感、电容等储能元件在某些接线方式下与电源频率发生谐振所造成的过电压。一般按起因分为：①线性谐振过电压。②铁磁谐振过电压。③参量谐振过电压。

（3）断电

常规生活中，我们把电源与用电设备之间脱离电器连接的现象称为断电。在电力系统中断电是如下描述的：在一定时间内,一相或多相完全失去电压(低于 0.1p.u.)称为断电。断电按持续时间分为三类: 瞬时断电 0.5 周波至 3s; 暂时断电 3s 至 60s； 持续断电>60s。

（4）与电压偏差的比较

表格 4 电压偏差与长时电压变化的比较

比较项	电压偏差	过电压和欠电压
针对环境	仅仅针对电力系统正常运行状态而言。	既可能出现在电力系统正常运行方式，也可能出现在电力系统非正常运行方式，如故障状态等。
偏差大小	电力系统在正常运行方式下，机组或负荷的投切所引起的系	过电压和欠电压强调实际电压严重偏离标称电压，分别为高于标

	统电压偏差并不大，其绝对值不大于标称电压的 10%。	称电压的 110 %和维持在标称电压的 10 %~ 90% ，
持续时间	电压偏差强调的是实际电压偏离系统标称电压的数值，与偏差持续的时间无关。	持续时间超过 1min

2.1.5 电压波动与闪变

（1）电压波动

电压波动（fluctuation）定义为电压均方根一系列相对快速变动或连续改变的现象，其变化周期大于工频周期。电压波动值为相邻电压方均根的两个极限值 U_{max}和 U_{min}之差△U，常以其标称电压的百分数表示其相对百分值:

$$d=\frac{U_{max}-U_N}{U_N}\times 100\% \qquad (2\text{-}2)$$

相对最大电压变动值:

$$d=\frac{U_{max}}{U_N}\times 100\% \qquad (2\text{-}3)$$

《GB12326-2000 电能质量-电压允许波动和闪变》中规定：在公共供电点的电压波动允许值如下：10kV 及以下为 2.5%，35kV—110kV 为 2%，220kV 及以上为 1.6%。

在配电系统运行中，这种电压波动现象有可能多次出现，变化过程可能是规则的、不规则的，亦或是随机的。变化幅值通常不超出 0.9～1.1 倍电压范围的一系列电压随机变化。在波动负荷中，以电弧炉引起的电压波动最为严重。多数国家在制定的电压波动与闪变标准中的条款是针对电弧炉负荷设定的。同时电弧炉造成的供电电压波动对用电设备和系统安全运行的影响主要决定于波动值的大小和变动的频度。

（2）闪变

闪变（flicker）是指电压波动对照明灯的视觉影响，是经过灯－脑－眼环节反映人对照度的主观视感。为更为本质地描述灯－脑－眼环节的频率特性，IEC 推荐引入视感度系数 K（f）：

$$K(f)=\frac{S=1觉察单位的8.8HZ正弦电压波动d\%}{S=1觉察单位的频率为f的正弦电压波动F\%}\times 100\% \qquad (2\text{-}4)$$

式中:S=1 觉察单位是以闪变觉察率 F(%)的 50%为瞬时闪变视感度的衡量单位，闪变觉察 F(%)的统计公式如下:

$$F(\%)=\frac{C+D}{A+B+C+D}\times 100\% \qquad (2\text{-}5)$$

式中：A 为没有觉察的人数；B 为略有觉察的人数；C 为有明显觉察的人数；D 为不能忍受的数。

电压闪变通常是以白炽灯的工况作为判断。闪变可分为周期性和非周期性两种，前者主要是由于周期性的电压波动引起的，如往复式压缩机、电弧炉等；后者往往与随机性电压波动有关，如电焊机等。影响闪变的其他因素还有照明装置、人的视感度等。通常人对照明变化需要有一定的视觉暂留时间， 高于或低于某一段频率的照明变化，普通人便察觉不到。在所有低频成份中，人眼对 8.8Hz 的电压波动最为敏感。因此，IEC 标准以 S＝1(S 为瞬时闪变视觉度)为察觉单位，对 1/2 以下的低频成分以 8.8Hz 为标准作归一化处理，通过 S=1 下的电压波动频率、电压波动及视觉系数之间的关系将不同频率下的低频电压波动转化为一个特定频率(如 8.8Hz)的电压波动，以该特定频率电压波动的限值作为判断是否发生闪变的标准，大于该限值则判断为发生了闪变；反之则没有，闪变与电压波动有着直接的关系,但由于引起闪变的某些量值难以量化，而且它还需要对电压波动(调幅波)频谱分析度进行统计。因此，对闪变的计算远远比计算电压波动要复杂得多。到目前为止还没有准确计算闪变的公式。

2.1.6 波形畸变

波形畸变，即实际电压波形对理想波形的偏差，可分为直流偏移、谐波、间谐波（interharmonics）、陷波（notching）和噪声等，其特征见表格 5。

表格 5 波形畸变特征

类别	频谱成分	电压幅值
直流偏移		0~0.1%
谐波	$0\sim100^{th}$	0~20%
间谐波	0~6KHz	0~2%

陷波		
噪声	宽带	0~1%

（1） 直流偏移

交流电压和电流中包含的直流分量，称之为直流偏移扰动。该扰动可以由电磁扰或半波整流引起。带来的危害是，致使变压器铁芯偏磁而达到饱和，导致变压器过热缩短使用寿命。

（2） 谐波

谐波(harmonics)是一个周期电气量的正弦波分量，其频率为基波频率的整数倍。谐波的国际公认定义是:“谐波是一个周期电气量的正弦波分量，其频率为基波频率的整数倍”。谐波的一个重要指标就是总谐波畸变率(THD)，定义为畸变波形因谐波引起的偏离正弦波形的程度。用公式表示为:

$$THD_h = \frac{\sqrt{\sum_{h=2}^{M} U_h^2}}{U_1} \times 100\% \qquad (2\text{-}5)$$

式中：THD_h－电压总谐波畸变率；U_h－各次谐波均方根值；U_1－基波均方根值；M－所考虑的谐波最高次数，由波形的畸变程度和分析的准确度要求来决定，通常取≤50。

我国《GB/T14549-93 电能质量-公用电网频率谐波》中规定：6—220kV 各级公用电网电压（相电压）总谐波畸变率是：0.38kV 为 5.0%，6—10kV 为 4.0%，35—66kV 为 3.0%，110kV 为 2.0%；用户注入电网的谐波电流允许值应保证各级电网谐波电压在限值范围内，所以国标规定各级电网谐波产生的电压总畸变率是：0.38kV 为 2.6%，6—10kV 为 2.2%，35—66kV 为 1.9%，110kV 为 1.5%。对 220kV 及其以上其供电的电力用户参照本标准 110kV 执行。

电力系统中的非线性负荷是造成波形畸变的主要根源。近年来，电力系统谐波问题日益严重。谐波会引起旋转电机的附加损耗、发热和振动，降低旋转电机的使用寿命。谐波电流在变压器绕组和线路传输中都要产生附加损耗，而系统变压器和线路的损耗构成了电网损耗的主要部分。在发生系统谐振或谐波放大的情况下，谐波的网损可达到相当大的程度(例如向电气铁道供电的电网在谐波严重

谐振时输变电网络中谐波损耗可达到电气铁道供电负荷的 1.85%)；谐波会对通信系统产生电磁干扰，降低电信质量，还会使重要的和敏感的自动控制、保护装置误动作。

产生原因是电力系统中某些设备的非线性特性或非线性负荷。带来的危害有：①引起旋转电机和变压器的附加损耗和发热，缩短其使用寿命；②谐波谐振过电压造成电气元件和设备故障；③对通信系统产生电磁干扰，影响通讯质量；④造成自动控制装置工作紊乱。

（3） 间谐波

间谐波(inter-harmonics) 是含有基波非整数倍频率的正弦电压或电流称，小于基波频率的分数次谐波也属于间谐波。间谐波一般来自电网中的静止式变频器、循环换流器、感应电机等。带来的影响是：干扰电力线载波通信，引起阴极射线管等显示装置的视感闪变。

（4） 陷波

陷波是一种周期性的电压扰动，主要由电力电子器件的电流换相引起。可通过受干扰电压的谐波频谱来表征。

（5） 噪声

噪声是指附加在电网电压/电流/中性线/信号线上的具有小于 200KHz 宽带频谱的非期望信号。噪声主要来自电力电子器件、控制电路、起弧设备和开关电源等。

2.1.7 三相不平衡

三相不平衡（unbalance）表现为电压的最大偏移与三相电压的平均值超过规定的标准，其数学定义为电力系统在正常的运行方式下，电量的负序分量均方根之与正序分量均方根值之比。

$$\varepsilon = \frac{U_2}{U_1} \times 100\% \qquad (2\text{-}6)$$

式中：U_1为三相电压正序分量的均方根值；U_2为三相电压负序分量的均方根值。

我国的《GB/T15543-1995 电能质量-三相允许不平衡度》中规定：电力系统公共连接点正常电压不平衡度允许值为 2%。短时不得超过 4%，标准中还规定对

每个用户电压不平衡度的一般限值为 1.3%。

电力系统的三相不平衡(或对称)是由于三相负载不平衡(或对称)以及系统元件参数的不对称所致。在研究不对称的三相电力系统时，广泛使用对称分量法，即将任何一组不对称的三相相量(电压或电流)分解成相序各不相同的三相对称的三相相量。三相电源电压畸变不对称时，对于三相四线制电路，电压中除含有谐波分量外，还含有正序、负序、零序分量。对于三相三线制电路，只含有正、负序分量。电力系统三相不平衡可以分为事故性不平衡和正常性不平衡两大类。事故性不平衡由系统中各种非对称性故障引起，比如单相接地短路、两相接地短路或两相相间短路等。而电力系统正常运行时，供电环节的不平衡或用电环节的不平衡都将导致电力系统的三相不平衡。

三相不平衡所产生的危害有：

① 对感应电动机: 电动机的负序电抗很小，所以负序电压产生的负序电流很大，使电动机的铜损增加。铜损的加大不仅使电动机效率降低，同时使电动机过热，导致绝缘老化过程加快。

② 对变压器：变压器处于不平衡负载下运行时，变压器容量得不到充分利用。研究表明，变压器工作在标称负载下，当电流不平衡度为 10%时，变压器绝缘寿命约缩短 16%。

③ 对换流器：三相不平衡使换流器的触发角不对称，换流器将产生较大的非特征谐波。非特征谐波电流的出现对换流器的谐波治理提出了更高的要求，直接导致换流器总投资的加大。

④ 对继电保护和自动装置：三相不平衡系统中的负序分量偏大，可能导致一些作用于负序电流的保护和自动装置误动作，威胁电力系统的安全运行，此外系统三相不平衡还会使某些负序启动元件对系统故障的灵敏度下降。

⑤ 对线路：在三相不平衡系统中，线路除正序电流产生的正序功率损耗外，还有负序电流及零序电流产生的附加功率损耗，因此加大了线路的总损耗，降低了电力系统运行的经济性。

⑥ 对计算机：在低压三相四线制系统中，三相不平衡引起中线上出现不平衡电流，中性点电位会产生漂移，严重时对计算机产生电噪声干扰，可能使计算机无法正常工作。

2.1.8 频率偏差

频率偏差(frequency deviation) 是指在电力系统正常运行条件下，系统频率的实际值与标称值(50Hz 或 6OHz，我国采用 5OHz 标准)之差。用公式表示为:

$$\delta f=f_{re}-f_N \qquad (2-7)$$

式中：δf－频率偏差，Hz；f_{re}－实际频率，Hz；f_N－系统标称频率，Hz。

对频率质量的要求全网相同，不因用户而异，各国对于该项偏差标准都有相关规定，包括互联电网和孤立电网中两种。我国电力系统的标准频率为 50Hz，《GB/T15945-1995 电能质量-电力系统频率允许偏差》中规定，电力系统正常频率偏差允许值为±0.2Hz，当系统的容量较小时，偏差值可以放宽到 0.5Hz，标准中没有说明系统容量大小的界限。在《全国供电规则》中规定“供电局供电频率允许的偏差”：电网容量在 300 万千瓦以上者为±0.2Hz，电网容量在 300 万千瓦以下者为±0.5Hz。实际的运行中，从各大电力系统看都保持在不大于±0.1Hz 范围内。

频率是电能质量的重要指标之一,系统负荷特别是发电厂厂用电负荷对频率的要求。要保证用户和发电厂的正常运行就必须严格控制系统频率，使系统的频率偏差控制在允许范围之内。允许频率偏差的大小不仅体现了电力系统运行管理水平的高低，同时反映了一个国家工业发达的程度。

当发电机与负荷出现有功功率不平衡时，系统频率就会产生变动，出现频率偏差。频率偏差的大小及其持续时间取决于负荷特性和发电机控制系统对负荷变化的响应能力。在任意时刻，系统中所有发电机的总输出有功功率如果大于系统负荷有功功率的总需求(包括电能传输环节的全部有功损耗)，那么，系统频率上升，频率偏差为正；反之，系统中所有发电机的总输出有功功率如果小于系统负荷对有功功率的总需求，系统频率则下降，频率偏差为负。电力系统的大事故，如大面积甩负荷、大容量发电设备退出运行等，会加剧电力系统有功功率的不平衡，使系统频率偏差超出允许的极限范围。系统有功功率不平衡是产生频率偏差的根本原因。

频率偏差过大所产生的危害有: ① 对用电负荷: 工业企业所使用的用电设备大多数是异步电动机，其转速与系统频率有关，系统频率变化将引起电动机转速改变，从而影响产品质量,降低劳动生产率。电动机的输出功率与系统频率有关，系统频率下降使电动机的输出功率降低，从而影响所传动机械的出力,使电

子设备不能正常工作，甚至停止运行,而电子设备对系统频率非常敏感，系统频率的不稳定会影响这些电子设备的工作特性，降低准确度，造成误差。② 对电力系统: 降低发电机组效率，严重时可能引发系统频率崩溃或电压崩溃。汽轮机在低频下运行时容易产生叶片共振，造成叶片疲劳损伤和断裂，处于低频率电力系统中的异步电动机和变压器其主磁通会增加，系统所需无功功率大为增加，导致系统电压水平降低，给系统电压调整带来困难。无功补偿用电容器的补偿容量与频率成正比，当系统频率下降时，电容器的无功出力成比例降低，不利于系统电压的调整，使感应式电能表的计量误差加大。

2.2 国际国内相关标准

电能质量一直为人们所关注[5,6,7]。不同的历史时期，其侧重点有所不同。

（1）稳态电能质量问题。早期，电网用电负荷一般由异步电动机、同步电动机、电热电炉、整流和照明设备等组成，其中异步电动机占的比例最大，电网中存在的电能质量问题主要有：电压偏差、谐波、三相不平衡以及电压波动和闪变等等，这类电能质量问题以波形畸变为主要特征，持续时间较长，故称之为稳态电能质量问题。

（2）动态电能质量问题的形成及其危害。自 20 世纪 80 年代以来，用电负荷结构发生了重大变化，诸如半导体整流器、晶闸管调压及变频调整装置、炼钢电弧炉、电气化铁路和新型家用电器等负荷迅速发展，由于其非线性、冲击性以及不平衡的用电特性，对电能质量造成了严重的“污染”，带来了一系列新的电能质量问题，如脉冲暂态（impulse）、电压暂升（swell）、电压暂降（sags）和瞬时供电中断（interrupt）等；与稳态电能质量问题相比，这类电能质量问题的发生更随机，持续时间更短，称之为动态电能质量问题。

电力系统大量应用的基于计算机、微处理器的自动控制设备，对电网中的动态电能质量问题更为敏感[8,9]。譬如，一个计算中心失去电压 2s 就可能破坏几十个小时的数据处理结果或者损失几十万美元的产值[10]；对于一个半导体芯片制造厂，甚至几分之一秒的电压暂降就可能使生产停顿、设备破坏或者出次品，造成百万美元级别的损失。

正因电能质量问题引发后果严重，各国电力组织及标准化组织纷纷制定相关

电能质量标准来定义这一问题。

1989 年，欧洲共同体决定制定电能质量的全面标准。1992 年 7 月欧洲电工标准化委员会(CENELEC)正式颁布《公用配电系统供电特性》文件(CENELECCLC/BT-TF68-6(sec)15)，作为欧洲共同市场对电能质量的统一标准，目前已为国际电工委员会(IEC)采用。

国际电工委员会(IEC)从电磁现象及相互干扰的方式考虑，给出了引起电磁干扰的基本现象分类：

1）传导型低频现象：谐波、间谐波、信号系统(电力线载波)、电压波动、电压凹陷(骤降)和间断、电压不对称、频率偏差、感应低频电压、交流电网中的直流分量；

2）辐射型低频现象：工频电磁场。

3）传导型高频现象：感应连续波电压或电流、单方向瞬变、振荡性瞬变。

4）辐射型高频现象：磁场、电场、电磁场、连续波、瞬变。

5）静电放电现象(ESD)。

6）核电磁脉冲(NEMP)。

美国电子电气工程师协会（IEEE）给出的关于电能质量领域电磁现象的具体分类如表格 6。表中给出了各种现象，并且描述了其属性和特征。电能质量问题主要分为稳态和非稳态两大类。对于稳态现象，可利用幅值、频率、频谱、调制、电源阻抗、陷落深度、陷落面积等属性进行描述；对于非稳态现象，可利用上升率、幅值、相位移、持续时间、频谱、频率、发生率、能量强度、电源阻抗等属性进行描述。相对于 IEC 标准中概念性的表述，表格 6 提供了一个更为清晰描述电能质量及电磁干扰现象的实用工具。

表格 6 IEEE 给出的电力系统电磁现象的特性及分类

类别			**典型频谱成分**	**典型持续时间**	**典型电压幅值**
暂态	脉冲暂态	纳秒级	5ns 上升沿	<50ns	
		微秒级	1μs 上升沿	50μs~1ms	
		毫秒级	0.1ms 上升沿	>1ms	
	振荡暂态	低频	<5KHz	0.3~50ms	0~4pu
		中频	5~500KHz	20us	0~8pu
		高频	0.5~5MHz	5us	0~4pu
短期变化	断电			0.5cycles~1min	<0.1pu

	电压骤降		0.5cycles~1min	0.1~0.9pu
	电压骤升		0.5cycles~1min	1.1~1.8pu
长期变化	持续停电		>1min	0.0pu
	欠电压		>1min	0.8~0.9 pu
	过电压		>1min	1.1~1.2 pu
电压不平衡			稳态	0.5~2%
波形畸变	直流偏移		稳态	0~0.1%
	谐波	$0\sim100^{th}$	稳态	0~20%
	间谐波	0~6KHz	稳态	0~2%
	陷波		稳态	
	噪声	宽带	稳态	0~1%
电压波动		<25Hz	断续	0.1~7%
电源频率变化			<10s	

我国国家质量技术监督局(原国家标准局)也将制定国家电能质量标准列为重点项目，至 2003 年底，已颁布了六项标准(其中二项标准已经修订过)即：

GB/T12325-2003《电能质量供电电压允许偏差》修订版

GB12326-2000《电能质量电压波动和闪变》修订版

GB/T14549-1993《电能质量公用电网谐波》

GB/T15543-1995《电能质量三相电压允许不平衡度》

GB/T15945-1995《电能质量电力系统频率允许偏差》

GB/T18481-2001《电能质量暂时过电压和瞬态过电压》

上述国家标准中，涉及的电能质量问题的允许限值如表格 7 所示。

表格 7 国家电能质量标准允许限值表

标准编号	**标准名称**	**允许限值**
GB 12325—1990	供电电压允许偏差	（1）35kV 及以上为正负偏差绝对值之和不超过 10%； （2）10kV 及以下三相供电为±7%； （3）220V 单相供电为+7%，-10%
GB/T 14549—1993	公用电网谐波	各级电网谐波电压限值

GB/T 15543—1995	三相供电电压允许不平衡度	（1）正常允许 2%，短时不超过 4%； （2）每个用户一般不得超过 1.3%
GB12326—2000	电压波动和闪变	电压变动 d 的限值和变动频度 r(h-1)有关：当 r≤1000 时，对于低压（LV）和中压（MV），d=1.25%－4%；对于高压（HV），d=1%－3%；对于随机不规则的变动，d=2%（LV、MV）和 d=1.5%（HV）10kV 及以下 2.5%，闪变限值如下表
GB/T15945—1995	电力系统频率允许偏差	（1）正常允许±0.2Hz，根据系统容量可以放宽到±0.5Hz； （2）用户冲击引起的频率变动一般不得超过±0.2Hz

随着国民经济和科学技术的飞速发展，我国电能市场逐步建立和完善，电网负荷也发生了较大的变化，使得现有的 6 项电能质量标准已不能满足实际需要。我国正在编制电能质量标准体系框架，其中暂态电能质量问题成为了目前工作的重点。

2.3 电能质量数据的监测

2.3.1 监测方式

通常的电能质量检测方式有：连续监测、定期或不定期监测和专门测量。

1)连续监测：一般用于重要变电站的公共供电点的监测。监测指标包括： 供电频率、电压偏差、三相电压不平衡度、谐波等。

2)定期或不定期监测：适用于需要掌握供电质量而不需要连续监测或不具备连续监测条件的监测方式。一般的公共供电点的供电质量通常只需定时监测。对于无冲击性负荷的电网，一般不存在明显的电压波动和闪变，这两个指标只需一二年测量一次即可。而对于有冲击性负荷的电网，往往一个月或一个季度测量一次，监测周期和每次的测量时间视具体情况而定。对于负荷稳定、用电量不大的电力用户， 电能质量指标比较稳定，则可一周或一个月定时监测一次。

3)专项测量：主要适用于干扰源设备接入电网(或容量变化)前后的电能质量

监测，用以确定电网电能质量指标的背景状况和干扰发生的实际量或验证技术措施效果。

2.3.2 监测设备

常见的电能质量监测设备常见的电能质量监测设备从简单到复杂， 大致可以分为三大类。

1）传统监测仪器：包括万用表、数字相机（ 记录扰动波形）、示波器（ 记录波形和时变数据）、扰动分析仪、谐波分析仪和频谱分析仪。但一般在需要时才对电能的质量进行检测，实时性差，当电能质量波动较大时，无法得到全面的质量信息。设备功能单一，一般只检测一二项电能质量指标。另外，它们大多安装在孤立的节点上，受器件和分析方法的限制，对系统中的短时暂态扰动难以快速、准确地捕捉，精度也往往达不到要求。

2）数字型监测仪器：采用单片机、数字信号处理器，一般都可和计算机相连， 构成数据处理能力较强的 PC + DSP 结构，用数值计算的方法对信号进行采集、解析与识别等加工处理，以达到提取信息和便于应用的目的：一方面改善监测速度和准确性，趋向于高性能的实时处理，例如数字式闪变测量仪；另一方面向多功能方向发展，例如扰动与谐波综合分析仪等。这类仪器对单个站点的测量有比较好的效果。不足之处在于：由于装置本身限制，无法同时监测多项指标；需要大量人力、物力进行测量、分析； 数据量有限，不利长期跟踪和深入评估。

3)智能型综合监测仪器：智能型电能质量监测仪的特点在于，除了对采集到的数据进行信号识别之外， 还可以进行进一步的分析处理， 从而提供更有意义的结论和建议。例如，扰动源的定位，应对措施的建议，报警功能以及各种专家系统。其他的优点将在后面结合电能质量监测系统进行说明。

选择仪器时有几项重要因素， 包括：

1 ）通道的数量(电压和/或电流)

2 ）仪器的温度要求和仪器的耐用性

3 ）输入电压范围和功率要求

4 ）测量三相电压的能力

5 ）输入隔离(输入通道之间和每个通道与地之间的隔离)

6 ）仪器装配(便携式、安装在机架上等)

7 ） 使用简便(用户接口、图形化功能等)

8 ） 通信能力(Modem，网络接口)

9 ） 分析软件的功能

10)仪器的综合性：单个仪器的功能越多，所需仪器数量越少。

2.4 电能质量数据的格式

随着大量电子类高科技产品的应用以及人们生活水平的日益提高，对于电网的电能质量提出了越来越高的要求；而另一方面，大量应用电力电子技术的工业设备和家用电器致使电网的电能质量逐渐恶化。例如大量变频“整流器”电弧炉等非线性负载的接入使得电网中的谐波污染情况日趋严重；超高压输电线路不循环换位和电动机车等大容量非对称负载的接入使局部电网的不对称度加大；大容量轧钢机等冲击性负载的接入造成电网的暂态干扰增大，电压闪变现象时常发生等等。

这些不仅会导致供用电设备本身的安全性降低,而且会严重削弱和干扰电网的经济运行，使得一些对电能质量要求严格的高科技产品无法使用,甚至于使它们产生错误的指令和控制动作，这在国防和一些关键行业会产生巨大的影响和后果，为此国家技术监督局相继颁布了涉及电能质量的国家标准。建立一个完善、高效、覆盖电网的电能质量管理信息系统已势在必行。

长期以来，我国的电能质量在线监测没有统一的标准， 一般都以设备厂家的监测仪器为核心，建立小范围的数据中心。通信主要基于 MODEM 和以太网等进行传输。这就造成了不同电能质量监测设备的后台分析软件不尽相同， 相互之间数据不兼容。通信和操作方式的不一致给信息交换带来很大的不便， 这在网络资源快速发展以及人们对信息资源要求共享的今天无疑是落后和不能被接受的。电能质量监测和分析涉及广泛的数据来源。如果多种类型的数据内容和描述格式各异，指标含义不统一， 必将导致电能质量监测和分析系统数据管理的混乱。

为了适应电能质量有关测量数据和计算数据的存储, 必须制定统一的数据存储体系, 作为数据采集、交换和分析的标准。因此, IEEE 标准委员会提出了一种电能质量数据的交换格式 PQDIF(Power Quality Data Interchange Format)。它完全独立于监测设备的软、硬件, 不仅可以较好地解决多数据源数据的兼容问题, 还

可以实现电能质量物理属性的多角度观察功能，满足了电能质量监测技术的发展需要。

电能质量数据交换格式 PQDIF 是一种平面文件结构。它由一系列逻辑相关的“记录”(Record) 链接而成，在每个记录中包含一系列元素，它定义了记录的内容。PQDIF 文件结构分为物理层和逻辑层。物理层描述文件的物理结构，使用标识识别文件的特定元素，它并不关心实际被储存的内容；逻辑层使用物理层定义的结构，利用特定标识在文件中建立元素，描述逻辑关系[11]。

2.4.1 PQDIF 的物理层结构

物理层结构就是指 PQDIF 的物理组成方式,它不涉及具体内容，由一系列相互关联的记录组成。这些关联由记录在文件中的绝对地址或者偏移地址组成，并全部记录在记录头中。这种结构相当于一个链表结构，每个记录头包括本身及其下一个记录在文件中的位置，相当于链表中的指针,可以很方便的对文件中的记录进行添加、插入和删除操作。

文件中的每一个记录都有相同的结构即都包含记录头和记录体。如表格 8 所示，记录头包括一个唯一标识符 GUID (Globally Unique Identifier) ,一个指定记录类型的标记符以及记录头大小、记录体大小等信息。

表格 8 记录结构说明表

	包含项目	举例
记录头	1.标识符：PQDIF 信息 2.标识：记录类型 3.记录头的大小 4.记录体的大小	8b235380-f29e-33de-7988-235987234987 标识信息包（tagContainer） 64byte 512byte
记录体	1.以一个集合开始 2.集合元素通过它在记录中的位置来标识，它与记录头中的标识是相对应的	集合 Colllection 集合元素数据 12 元素 0 标识符：文件名 类型：向量数据或等级值 物理类型 内容

记录体由一系列的元素组成。PQDIF 共有 3 种类型的元素: 集合(Collection)、等级值(Scalar)和向量数组(Vector)。集合由标识符和与其他元素的相对链接组成;等级值是指一个元素的物理类型，如 INT4 (4 字节整数) , CHAR1 (1 字节字符型) ;向量数组是一个指定了物理类型的且可变大小的数组，即元素的内容。

2.4.2 PQDIF 的逻辑层结构

逻辑层结构由一系列记录组成, 共有 4 种类型的记录: 信息包记录、数据源记录、监控器设置记录和观察值记录。信息包记录比较特殊, 它记录了文件的结构信息, 位于文件的开始部分, 更像是一个查询目录, 在一个文件中有且只有一个信息包记录。而其它 3 种记录在一个文件中则可多可少, 也可以没有。文件的基本结构就是由各种记录的逻辑分层组成, 即由一个信息包记录联接着一个或多个数据源记录、监控器设置记录、观察数据记录组成, 如图 3 所示。

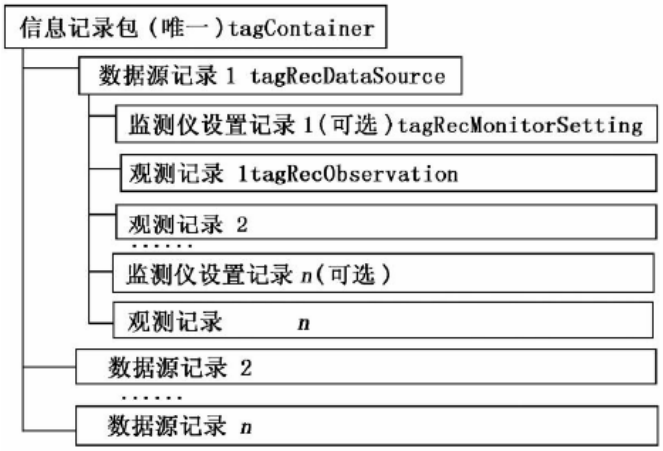


图 3 PQDIF 物理结构示意图

2.5 数据的采集、转换与存储

2.5.1 电能质量监测系统与数据采集

电能质量监测系统的初级形式是分布式监测装置，可以实时监测、分析一个变电站多条出线的状态，进而利用网络技术，逐步发展为分布式监测系统。Internet 和通信技术的发展，为信息共享和数据交换提供了便利，也为电能质量监测的网络化创造了条件。网络化的电能质量监测系统也在不断发展完善。GPS 技术被引入到电能质量监测系统中，用以保证采样数据的同步性和准确性。设计了专门的电能质量网络监测仪的上网接口，改善了联网传输性能。有人提出一种基于互联网地理信息系统(WebGIS)的电能质量监测系统，将电能质量信息与地理分布信息结合起来提供给电网管理人员，且用户终端只需一个支持 Java 编程语言的 Web 浏览器就可以完成监测的基本操作，降低了系统维护和管理费用。虚拟仪器技术也应用到电能质量监测系统中。也有人以虚拟仪器技术为平台构建了电能质量网络化监测系统，通过 DataSocket 技术实现了对监测数据的动态传输。

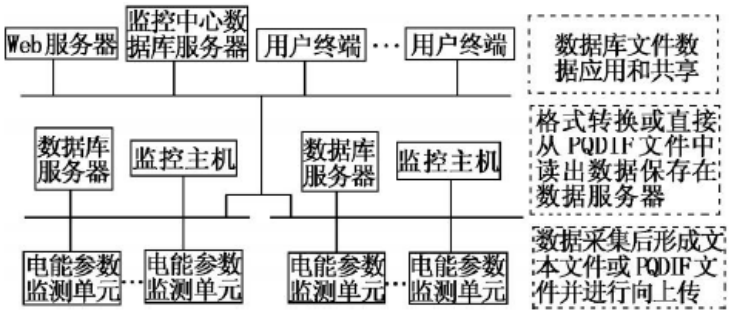


图 4 电能质量管理系统整体结构

电能质量管理系统的整体结构如图 4 所示，系统采用 3 层分布式网络结构。第 1 层为处于变电站(所) 的电能质量监测设备，用于收集电能参数并进行简单的分析形成文本文件或者直接以 PQDIF 文件格式进行保存，并负责上传至监控主机。一些报警报文必须实时上传并作相应联动，而测量量按照实时等级进行分时上传。第 2 层为监控主机和数据库服务器，监控主机负责接收来自监测设备的数据，并从 PQDIF 文件读出数据放入数据库服务器，终端用户共享数据库信息。第 3 层为管理层，包括后台数据库和 Web 应用发布系统，由服务器、管理终端、用户终端等组成。监控主机通过局域网与下层的监测设备进行通信，各监测设备可以主动上传数据给监控主机，监控主机也可以通过召唤传输任一监测设备的数据。监测设备的接入可以采用以太网或者 MODEM 等方式接入。监控主机与监测设备的通信规约采用统一的控制格式。

监测设备主要对有关电能质量参数的电压和电流有效值实时监测及超标报警，同时对电压合格率(电压偏差)、三相不平衡度、电压波形畸变率及闪变等数据监测与采集，监测设备还需要具备通信、输入输出、显示以及后台统计数据存储和分析功能。这些参数与功能的实现参照国家标准或行业规范，数据存储与通信则采用 PQDIF 标准。

监控主机相当于一个中间数据源，起到承上启下的作用。对下除了有限的控制外主要是完成与下位机的数据文件采集与格式转换工作；对上则是将数据进行二次处理后报送到指定的后台数据库。

监控主机内有数据处理和数据缓存模块，可以在本级生成电能质量数据分析报告和查询短期历史数据。后台数据库和 Web 应用发布系统是一种基于服务器/客户端的网络构架。本级提供的数据是一个地区电网的电能质量报表和数据。既可以反映本地区电能质量的长期运行趋势走向，也可以反映本地区在某一段时间内的电网运行状况。这对于供电企业了解和掌握电网运行参数、建立切实可行

的管理计划提供了有力参考。在后台数据库的基础上可以建立综合信息查询系统,建立能对电能质量数据进行统计分析处理、评估和预测的专家系统以及基于国标的规则库，能分区域、分时段、分电压等级、分性质对电能质量指标进行分析比较、趋势比较等，强大的数据处理功能是该类系统的一个特点[12]

现代电网规模越来越大，监测点越来越多，未来电能质量的监测要实现不同供电点甚至多个供电系统的集中监测。在功能上，更强调智能化，除具有计算、显示功能外，还要有一定的判断、决策功能，例如能进行事件预测、故障辨识、干扰源识别和实时控制等，初步具有自动的、实用先进的智能评估功能。电能质量在实现了在线监测、实时分析的基础上，正向着网络化、信息化、标准化和智能化的方向发展完善。网络化、信息化、标准化和智能化已成为电能质量监测系统的必然发展趋势， 它为电网的优化和事故分析提供实时可靠的数据， 为电能质量综合评估提供切实依据，也是电力企业面向市场，适应竞争的强有力手段，可以进一步保障各级用户的正常用电秩序，为其提供优质的电能。

2.5.2 PQDIF 数据的压缩与转换

电能质量监测系统产生的原始数据通过 PQDIF 与应用层的数据库打交道，因此, PQD IF 是比数据库更低层的数据源信息。它一方面屏蔽了数据库面对各种不同监测设备的数据复杂性,同时通过 PQDIF 这一通用平台向下协调了异构数据源,向上为访问数据的应用层提供了统一的数据格式和访问接口。这不仅可以简化、统一应用程序开发人员的开发环境，减少程序设计的复杂性，还便于汇集不同数据源上的数据[13]。

外监测仪器作为终端设备一般不可能设计成大容量存储设备，故而这种方法在实际工程中应用较少。其二是厂家选择习惯的方法储存数据，再使用特定的程序将监测设备中原来的数据文件转换成 PQDIF 文件。工程实现中监测设备将获得的原始数据以文本格式进行存储，在进行通信时由监控主机完成文件格式的转换工作。

应用中的 PQDIF 都是经过压缩的，所以有必要对其压缩方法有所了解。PQDIF 文件采用没有版权限制的 ZLIB 库，运用 LZ77 算法进行压缩。文件在压缩时也具有特殊性，它采用一种叫做记录级的压缩方法，即所有的记录头都不压缩，而且文件的第一个记录——信息包记录也不压缩。这种压缩方法使 PQDIF 文件在被应

用程序读取时不必对文件的条目内容进行阅读和解压缩就能够很快地了解文件的结构。

PQDIF 的转换一般有 2 种方式。其一是由电能质量监测设备直接进行转换，即在前置单元和监控主机通信前完成格式转换，直接以 PQDIF 格式存储数据。这种方法不仅要求监测设备本身具有较大的存储空间，而且要求监测设备具有较高的 CPU 处理能力。因为在实际工程应用中，高质量监测设备常采用处理运算量较大的算法程序(如 FFT、小波算法) ,而且监测参数的获得实时性要求较高，若再进行 PQDIF 格式的转换会大幅提高 CPU 的使用率；另外监测仪器作为终端设备一般不可能设计成大容量存储设备，故而这种方法在实际工程中应用较少。其二是厂家选择习惯的方法储存数据，再使用特定的程序将监测设备中原来的数据文件转换成 PQDIF 文件。工程实现中监测设备将获得的原始数据以文本格式进行存储，在进行通信时由监控主机完成文件格式的转换工作。

2.5.3 电能质量数据存储实例

当电能质量数据被从监测节点以 PQDIF 的格式传输到省一级的数据中心（同时也是电能质量监测管理中心）后，通常这些数据被按内容含义解析后以存储在关系型数据库之中。

以某省电网的电能质量监测系统为例，该中心数据库有采集全省 1000 多个监测点的电能质量数据。这些数据存储在 DB2 关系数据库中，按数据种类、用途组织成几十个关系型数据表。主要有如下几类：

1）数据采集情况记录：数据采集统计表、采集工况明细表；

2）用户情况表：用户关联线路、用户历史数据关联线路、用户治理设备表及历史表、用户表及历史表、用户负荷及历史表；

3）暂态数据记录表：历史数据超标表、录波数据、骤变数据、骤变记录；

4）测试报告产生的电压谐波、电流谐波、谐波功率、非谐波电能质量指标等表；

5）历史数据记录表：

3 秒的功率表、3 秒电压数据表、3 秒电流数据表、3 秒闪变数据表；

历史数据功率表及无功功率扩展、有功功率扩展、视在功率扩展；

历史数据电压表及其扩展表、历史数据电流表及其扩展表；

汇总日数据功率表及无功功率扩展、有功功率扩展、视在功率扩展；

汇总日数据电压表及其扩展表、电流表及其扩展表；

汇总日闪变表，及闪变表；

6）历史数据分周期统计表：

95 超标统计表

功率动态统计表及无功功率扩展、有功功率扩展、视在功率扩展；

日统计数据：功率及无功功率扩展、有功功率扩展、视在功率扩展、电压表及其扩展表、电流表及其扩展表、超标、闪变；

周统计数据：功率及无功功率扩展、有功功率扩展、视在功率扩展、电压表及其扩展表、电流表及其扩展表、超标、闪变；

月统计数据：功率及无功功率扩展、有功功率扩展、视在功率扩展、电压表及其扩展表、电流表及其扩展表、超标、闪变；

季度统计数据：功率及无功功率扩展、有功功率扩展、视在功率扩展、电压表及其扩展表、电流表及其扩展表、超标、闪变；

年统计数据：功率及无功功率扩展、有功功率扩展、视在功率扩展、电压表及其扩展表、电流表及其扩展表、超标、闪变；

电网侧电能质量指标统计汇总表，包括省级、市级、县级、监测点各级按月度、季度、年度统计汇总表；

非线性客户电能质量汇总表，包括省级、市级、县级、监测点各级按月度、季度、年度统计汇总表，以及非线性客户信息统计报表；

动态统计表：电压动态统计表及其扩展表、电流动态统计表及其扩展表、监测点动态统计、超标动态统计表、闪变动态统计表；

数据合并日志表、日期代码表。

7）预警相关表：安全预警结果表、母线评估结果、线路评估结果、预警线路表、母线综合标记位、线路综合标记位、电压指标限值表、稳态指标预警阀值表、预警流程电压暂降预警阀值、稳态指标异常事件记录表、电压暂降信息记录表、回溯预警时间临时表、执行过预警的日期列表、预警任务表。

8）电网基本信息表：单位表、变电所关联、变电站（PMIS）（监测网）、数据采集工况、母线、母线关联、电压通道表、限值表、电流容量维护表、电流通道表、限值表、离线设备表及使用记录表、线路表、线路关联表、组织架构、设

备信息、设备维护表。

9）评估相关表：县级及市级公司评估结果表、变电站评估结果表、回溯评估月份表、已评估月份记录表、电压及电流通道评估任务表。

以其中记录电能质量主要问题的电能质量历史数据超标表（dat_overrun）为例，其基本数据结构如下图所示。

OID	CHV_OID	CHI_OID	DAT_OID	KIND	OC_DATE	SEQ	COL
1304382424	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"11"
1304382425	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"24"
1304382426	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"25"
1304382427	0	3131	6898101203	6	"2012-08-15 00:20:00"	"A"	"24"
1304382428	0	3131	6898101203	6	"2012-08-15 00:20:00"	"A"	"25"
1304382429	0	3131	6898101204	6	"2012-08-15 00:30:00"	"A"	"24"
1304382430	0	3131	6898101204	6	"2012-08-15 00:30:00"	"A"	"25"
1304382431	0	3131	6898101205	6	"2012-08-15 00:40:00"	"A"	"24"
1304382432	0	3131	6898101205	6	"2012-08-15 00:40:00"	"A"	"25"
1304382433	0	3131	6898101206	6	"2012-08-15 00:50:00"	"A"	"24"
1304382434	0	3131	6898101206	6	"2012-08-15 00:50:00"	"A"	"25"

图 5 电能质量历史数据超标表结构示意图

其中，chi_oid 是线路监测点 ID，kind 代表电能质量问题类型，其中：1 代表频率，2 代表短闪变，3 代表长闪变，4 代表电压不平衡，5 代表谐波电压，6 代表谐波电流，7 代表总谐波畸变率，8 代表电压变动，9 代表电压偏差。

oc_date 是记录时间，seq 代表出问题的相位，col 字段用于记录谐波波次。

其它数据表类似地以时间为纵轴，以节点 ID 为区分记录着各监测节点采集的电能质量数据以及衍生出的相关统计数据，在此不一一列举。

3 电能质量高级分析概述

电能质量监测系统为电力系统实时地积累了海量的监测数据。相对于面向基本目的的电能质量数据初级分析，电能质量数据高级分析的“高级”体现在以下5个方面：

1） 在分析目的上，由基本分析的面向统计目标转向面向预测目标；

2） 在分析目的上，从面向已经已知问题转向面向未知规律的研究；

3） 在分析范围上，从面向单个监测节点数据的分析转向面向某个区域网络的分析；

4） 在分析手段上，从单点计算为主转向面向海量数据的分布式分析技术；

5） 在分析时限上，从滞后的批量数据分析转向开始关注实时分析数据的需求。

上述高级分析可以通过对数据挖掘及其扩展技术的深化应用来实现。

3.1 从统计到预测

基于电能质量监测系统采集来的数据，我们可直观地想到对各地区各监测点按各类电能质量问题进行统计，基于统计结果可以评估各地区的电能质量状况。

我们还希望能够从统计结果中发现存在的问题以及发生问题的原因，或者在原因不易查明的情况下，我们能够预知电网中电能质量问题的未来发展趋势，这些就要求我们不能仅满足于情况统计，而应进一步进行电能质量相关的预测分析。

在预测目标上，有如下几个比较明确的方向：

1）基于预测进行电能质量扰动源识别；

2）基于预测对电能质量整体水平进行预测；

3）基于预测进行电能质量改进；

目前对于电能质量预测的研究和应用还相对较少，但是无疑地，任何能够准确预测未来发展变化的高级分析都将极大促进电能质量服务水平的提高。

3.2 从已知到未知

电能质量问题是在电力系统运行过程中产生的，针对电力系统的传统分析方法主要是掌握电力线路的基本结构和主要参数，并对电力线路、变压器、发电机与负荷等要素及整个电力网络进行数学建模，在此基础上进行相关的潮流计算、

对称短路或不对称短路的分析计算、静态稳定性分析、暂态稳定性分析、电力系统振荡分析等计算，以此掌握系统运行情况以及其中蕴藏的问题。

上述计算需要确切知道电力系统每条线路上的具体参数，并明确知道节点间的位置关系及相互影响。对电力系统持续的电能质量监测积累了大量的数据，我们希望从这些数据中发现一些我们没有考虑到关联或规律，我们还希望能够通过对数据的观察和分析，找到电力系统在电能质量问题上的分布趋势或体现出的不同特征。

3.3 从单点到区域

在大多数情况下，需要对单个监测点的数据进行分析，以了解该点的电能质量分布情况和变化趋势，或者找出暂态电能质量事件的原因。就单个监测点而言!在某一时刻，描述电能质量的各指标之间存在确定的相互关系，但是此时通常仅有一个或少数几个指标是主要的，根据这几个主要的指标就可以比较清晰地分析出该监测点的电能质量信息。例如由单相接地引起的电压暂降、暂升和短时中断等暂态电能质量问题符合单相接地的基本特征，基于这些特征，分析人员就容易从离散的指标中分析到其统一的物理现象，从而抓住关键，给出判断。

但是，对同一时刻的物理现象所表现的指标进行分析，只能看出相关指标的微观联系，通常用于针对暂态电能质量问题的分析，以找出暂态电能质量事件的原因。对于稳态电能质量指标而言，这种微观的分析方法是不太适用的，因为往往某个时刻的一组稳态电能质量指标没有多大的实用意义。现实中往往结合一段时间内的统计数据进行稳态电能质量指标的统计分析，各指标的统计结果一般不反映诸多现象的时刻特征，但是能够反映其可能包含的物理现象的综合特征模型，便于进一步仿真分析，并提出综合的控制措施。

随着电能质量监测点的增多，监测网络的扩大，管理人员在时间和精力上往往不能够定期对每个监测点都进行详细的、独立的电能质量分析。

考虑到电网的特征，电能质量的各个指标实际上在整个电网内有一个传递、相互影响的过程。 如谐波在整个电网内的传播，以及某个监测点的电压暂降对其相邻线路的影响等。 因此，针对电能质量指标的分析，必将会发展到一个系统性的层次上，即在整个电网结构的层面上进行全局的、系统的电能质量评估和分析。当电能质量监测网络达到一定规模后，系统所面对的最大问题就是海量数

据的问题。海量数据会在系统存储容量、查询性能等方面给电能质量监测系统带来很大考验，甚至可能导致整个系统的崩溃。采用系统性的分析方法，最大的优势在于可以减少直接面对无用的冗余数据，而只需要面对真正需要深入分析的电能质量问题。这种情况下，可以利用数据挖掘技术来实现系统性电能质量指标分析。采用数据挖掘技术以后，电能质量监测系统呈现给使用者的是经过统计、整理出来的数据，通过这些数据可以很清晰地了解整个电网内各个监测点的电能质量总体情况，然后可以选择对于电能质量问题比较严重的监测点进行深入的分析。此时可以采用针对单个监测点的宏观或者微观的分析方法。

3.4 从集中到分布

电能质量监测数据来成区域的电力系统的监测，每天汇集大量的数据，对这些数据进行分析已经不能满足于传统上的集中到一台机器上的分析。一方面，由于汇集的数据已经开始采用分布式存储，同一时期的电能质量监测数据可能分散存储在多台计算机上；另一方面，对电能质量数据进行高级分析所采用的算法通常要需要大量的 CPU 时间和内存。传统上很多基于一台主机进行计算的算法已经不适用于新的数据环境下的需求。

因而，面向分布式存储的电能质量数据和高计算复杂度（包括时间复杂度和空间复杂度）要求的高级分析，电能质量高级分析的技术体系需要首先考虑基于分布式计算思想进行架构。

3.5 从批量到实时

电能质量监测数据是对运行中的电力系统进行实时监测而产生的，同时，电力系统中的各条线路及各个设备都是在环境中不断发生变化的。传统的电能质量数据分析通常是

行高级分析所采用的算法通常要需要大量的 CPU 时间和内存。传统上很多基于一台主机进行计算的算法已经不适用于新的数据环境下的需求。

因而，面向分布式存储的电能质量数据和高计算复杂度（包括时间复杂度和空间复杂度）要求的高级分析，电能质量高级分析的技术体系需要首先考虑基于分布式计算思想进行架构。

4 数据挖掘技术

4.1 基本定义与挖掘任务

数据挖掘（Data Mining，DM）又称数据库中的知识发现（Knowledge Discover in Database，KDD），是目前人工智能和数据库领域研究的热点问题，所谓数据挖掘是指从数据库的大量数据中揭示出隐含的、先前未知的并有潜在价值的信息的非平凡过程。数据挖掘是一种决策支持过程，它主要基于人工智能、机器学习、模式识别、统计学、数据库、可视化技术等，高度自动化地分析企业的数据， 做出归纳性的推理，从中挖掘出潜在的模式，帮助决策者调整市场策略，减少风险，做出正确的决策。

知识发现过程由以下三个阶段组成：（1）数据准备，（2）数据挖掘，（3）结果表达和解释。 **数据准备**是从相关的数据源中选取所需的数据并整合成用于数据挖掘的数据集；**规律寻找**是用某种方法将数据集所含的规律找出来；**规律表示**是尽可能以用户可理解的方式（如可视化）将找出的规律表示出来。

数据挖掘可以与用户或知识库交互。并非所有的信息发现任务都被视为数据挖掘。例如，使用数据库管理系统查找个别的记录，或通过因特网的搜索引擎查找特定的 Web 页面，则是信息检索（information retrieval）领域的任务。虽然这些任务是重要的，可能涉及使用复杂的算法和数据结构，但是它们主要依赖传统的计算机科学技术和数据的明显特征来创建索引结构，从而有效地组织和检索信息。尽管如此，数据挖掘技术也已用来增强信息检索系统的能力。数据挖掘面向广泛的应用领域，不同的挖掘目的和数据形式决定了数据挖掘任务的多样性。数据挖掘任务用于指定数据挖掘要找的模式类型。一般而言，数据挖掘任务可以分为两类：描述和预测。描述性挖掘任务描述数据集合中数据的一般性质；预测性挖掘任务对当前数据进行推断，以做出预测。

根据目标模式的不同，数据挖掘任务主要可以分为：概念/类描述、频繁模式挖掘、分类与预测、聚类分析、离群点分析和演变分析等几类。

（1） 概念/类描述

数据可以与类或概念相关联。用汇总的、简洁的和精确的方式描述、各个类和概念可能是有用的。这种类或概念的描述称为概念/类描述（Class/concept describtion）。这种描述可以通过下述方式得到：1）数据特征化，一般地汇总

所研究类的数据；2）数据区分，将目标类与一个或多个可比较类进行比较；3）数据特征化的比较。

（2） 频繁模式挖掘

频繁模式（frequent pattern）是在数据中频繁出现的模式。存在多种类型的频繁模式，包括项集、子序列和子结构。通常，频繁项集是指频繁地事务数据集中一起出现的项的集合；频繁子序列是针对同一识别标识（如：用户 ID）下所产生事务按时序上的先后关系经常出现的序列模式;子结构可能涉及不同的结构形式，如图、树或格，如果一个子结构频繁出现，则称它为（频繁）结构模式。

对频繁模式的挖掘称为关联分析，分析结果是一系列的关联规则。通常关联规则是那些同时满足最小支持度阈值和最小置信度阈值的，而不能同时满足则被认为是不令人感兴趣的而被丢弃。根据不同要求，还可以做相关联的属性-值对之间的其它相关分析（如：提升度、全置信度、χ^2度量、余弦度量）。

（3） 分类与预测

分类（Classification）是找出能够描述和区分数据类或概念的模型（或函数），用于预测类标号未知的对象类。导出模型是基于对训练数据集（即类标号已知的数据对象）的分析。

预测（Predication）是建立连续值函数模型，用于预测空缺的或不知道的数值数据，而不是类标号。驾照分析（regression analysis）是一种最常用的数值预测统计学方法。预测也包含基于可用数据的分布趋势识别。

相关分析（relevance analysis）可能需要在分类和预测之前进行，它试图识别对于分类或预测过程无用的属性，这些属性应当排除。

（4） 聚类分析

聚类（Clustering）分析数据对象不考虑已知的类标号。一般情况下，训练数据中不提供类标号，因为开始并不知道类标号。可以使用聚类产生这种标号。对象根据最大化类内部的相似性、最小化类之间的相似性的原则进行聚类或分组。也就是说，对象的簇（Cluster）形成簇内对象具有很高的相似性，面与其它簇中对象很不相似。所形成的簇可看作一个对象类，由它可导出规则。

（5） 离群点分析

数据集中可能包含一些数据对象，与数据集一般行为或模型不一致。这些数据对象被称为离群点（Outlier）。大部分数据挖掘方法将离群视为噪声或异常

而丢弃，而有些应用（如欺骗检测）中，罕见的事件可以比正常出现的事件更令人感兴趣，因而展开离群点挖掘（outlier mining）

可以假定一个数据分布或概念模型，使用统计检验检测离群点；或者使用距离度量，将远离任何簇的对象视为离群点。基于偏差的方法通过考察一群对象主要特征上的差别来识别离群点，而不是使用统计或距离度量。

（6） 演变分析

演变分析(Evolution analysis)描述行为随时间变化的对象的规律或趋势，并对其建模。尽管这可能包括时间相关数据的特征化、区分、关联和相关分析、分类、预测或聚类。这类分析的不同特点包括时间序列数据分析、序列或周期模式匹配和基于相似性的数据分析。

4.2 典型算法及特点

国际权威的学术组织 the IEEE International Conference on Data Mining (ICDM) 2006 年 12 月评选出了数据挖掘领域的十大经典算法：C4.5， k-Means，SVM， Apriori， EM，PageRank，AdaBoost，kNN，Naive Bayes，和 CART。

不仅仅是选中的十大算法，其实参加评选的 18 种算法，实际上随便拿出一种来都可以称得上是经典算法，它们在数据挖掘领域都产生了极为深远的影响。

（1） C4.5

C4.5 算法是机器学习算法中的一种分类决策树算法,其核心算法是 ID3 算法. C4.5 算法继承了 ID3 算法的优点，并在以下几方面对 ID3 算法进行了改进：

1) 用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足；

2) 在树构造过程中进行剪枝；

3) 能够完成对连续属性的离散化处理；

4) 能够对不完整数据进行处理。

C4.5 算法有如下优点：产生的分类规则易于理解，准确率较高。其缺点是：在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，因而导致算法的低效。

（2） K-Means

k-means 算法是一个聚类算法，把 n 的对象根据他们的属性分为 k 个分割，k < n。它与处理混合正态分布的最大期望算法很相似，因为他们都试图找到数据中自然聚类的中心。它假设对象属性来自于空间向量，并且目标是使各个群组内部的均 方误差总和最小。

（3）SVM

支持向量机（Support Vector Machine，简称 SV 机，论文中一般简称 SVM）。它是一种监督式学习的方法，它广泛的应用于统计分类以及回归分析中。支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的超平面的两边建有两个互相平行的超平面。分隔超平面使两个平行超平面的距离最大化。假 定平行超平面间的距离或差距越大，分类器的总误差越小。一个极好的指南是 C.J.C Burges 的《模式识别支持向量机指南》[14]。van der Walt 和 Barnard 将支持向量机和其他分类器进行了比较。

（4） Apriori

Apriori 算法是一种最有影响的挖掘布尔关联规则频繁项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则。在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。

（5）EM

在统计计算中，最大期望 （EM，Expectation－Maximization）算法是在概率（probabilistic）模型中寻找参数最大似然 估计的算法，其中概率模型依赖于无法观测的隐藏变量（Latent Variabl）。最大期望经常用在机器学习和计算机视觉的数据集聚（Data Clustering）领域。

（6）PageRank

PageRank 是 Google 算法的重要内容。2001 年 9 月被授予美国专利，专利人是 Google 创始人之一拉里·佩奇（Larry Page）。因此，PageRank 里的 page 不是指网页，而是指佩奇，即这个等级方法是以佩奇来命名的。

PageRank 根据网站的外部链接和内部链接的数 量和质量俩衡量网站的价值。PageRank 背后的概念是，每个到页面的链接都是对该页面的一次投票， 被链接的越多，就意味着被其他网站投票越多。这个就是所谓的“链接流行度”——衡量多少人愿意将他们的网站和你的网站挂钩。PageRank 这个概念引自 学术中一

篇论文的被引述的频度——即被别人引述的次数越多，一般判断这篇论文的权威性就越高。

（7）AdaBoost

Adaboost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器)。其算法本身是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器最后融合起来，作为最后的决策分类器。

（8） kNN

K 最近邻(k-Nearest Neighbor，KNN)分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别。

（9）Naive Bayes

在众多的分类模型中，应用最为广泛的两种分类模型是决策树模型(Decision Tree Model)和**朴素贝叶斯模型**（Naive Bayesian Model，NBC）。朴素贝叶斯模型发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。同时，NBC 模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。理论上，NBC 模型与其他分类方法相比具有最小的误差率。 但是实际上并非总是如此，这是因为 NBC 模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的，这给 NBC 模型的正确分类带来了一定影响。在属 性个数比较多或者属性之间相关性较大时，NBC 模型的分类效率比不上决策树模型。而在属性相关性较小时，NBC 模型的性能最为良好。

（10）CART

分类与回归树(Classification and Regression Trees，CART,)有两个关键的思想：第一个是关于递归地划分自变量空间的想法；第二个想法是用验证数据进行剪枝。

其它更多、更详细的算法说明将在后续章节具体的电能质量应用中阐述。

数据挖掘技术本质上是一系列对数据集合进行形式转换处理的算法，而这些算法存在的目的是从数据集合中发现“新的、潜在的、有价值的”的信息。

“新的”是指以前未被发现的，“潜在的”是指要发现的信息并不是显而易见的，而“有价值的”则是相对于挖掘目的而言，是针对领域问题的。概括来说，挖掘目的主要有两种：一是“描述”，通过对原始数据集合（通常是一个二维表）的形式变换，形成多个、新形式的子集合，从而看清数据集合在数据含义上的分化关系，发现相近数据关系和离群数据(如分类、聚类、离群点分析及数据可视化)；二是“预测”，通过对原始数据的变换处理，使数据中所蕴含的在时序上的变化规律或在数据间的横向关联规律以规则的方式记录，从而为对新数据的属性预测提供依据（如：预测、频繁序列挖掘、关联分析和演变分析）。

数据挖掘利用了来自如下一些领域的思想：(1) 来自统计学的抽样、估计和假设检验，(2) 人工智能、模式识 别和机器学习的搜索算法、建模技术和学习理论，(3)数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。一些其他领域也起到重要的支撑作用。特别地，需要数据库系统提供有效的存储、索引和查询处理支持。源于高性能（并行）计算的技术在处理海量数据集方面常常是重要的。分布式技术也能帮助处理海量数据，并且当数据不能集中到一起处理时更是至关重要。

从机器学习的角度，数据挖掘的方法可以分为两大类：一类是基于有监督学习的数据挖掘方法，另一类是基于非监督学习的数据挖掘方法。

基于有监督学习的数据挖掘方法是指利用可用的数据建立一个模型，这个模型对剩余的数据中某个特定的变量进行描述。其基本过程如下图所示：

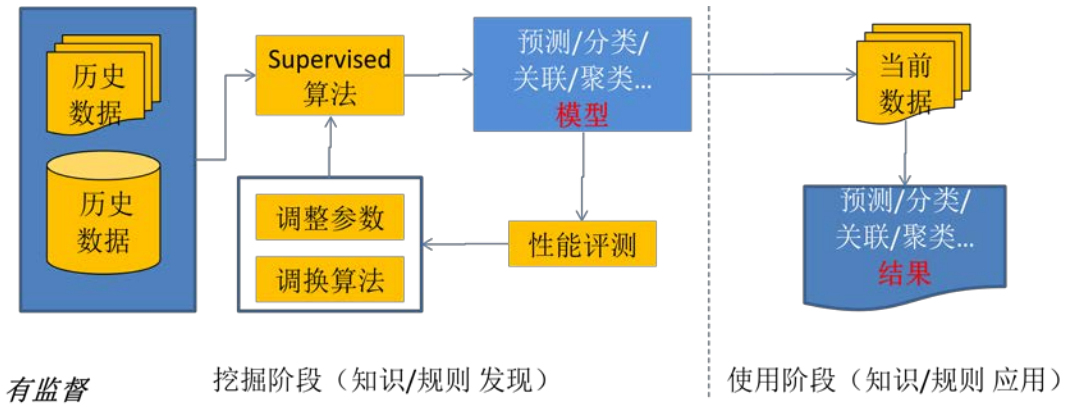


图 6 有监督学习类挖掘算法流程

基于非监督学习的数据挖掘方法是指没有从目标数据中选出某一具体的变量用模型进行描述，而是在所有的变量中建立起某种关系。其基本过程如下图所示：

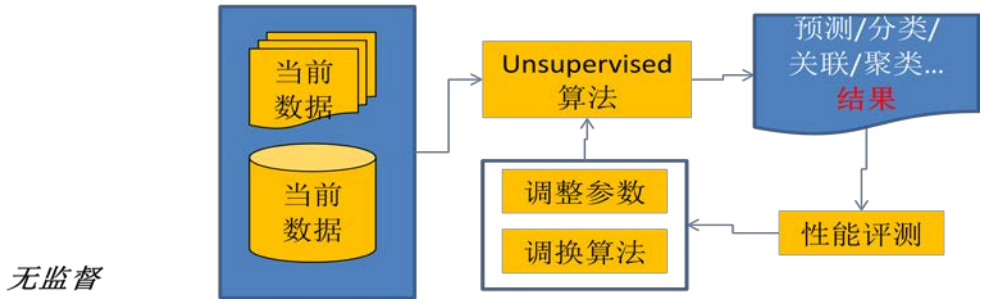


图 7 无监督学习类挖掘算法流程

在上述数据挖掘任务中，分类与预测通常属于有监督学习方法；频繁模式挖掘、聚类分析、离群点分析通常是属于非监督学习方法；概念/类描述分阶段属于两类方法；演变分析则既包含有监督学习方法也包含非监督学习方法。

针对电能质量数据的挖掘分析,要从已有的电能质量数据中发现数据之间在电能质量衡量含义上的关系和变化规律，从而制定相应的规则，为优化或控制电能质量提供依据。

其它算法（如:数值计算、逻辑推理、统计、数据形式转换等）的特点是：利用规则处理数据，规则不依赖于数据而存在。

而数据挖掘算法（如：预测、分类、聚类、关联、离群点、演变）的特点是：利用数据产生规则，**规则的内容依赖于数据的含义及特点**。

4.3 挖掘对象与应用原则

（1）挖掘对象

从处理对象来看，挖掘的数据包括结构化、半结构化和非结构化数据。**结构化数据**主要是最小单元为简单类型（字符、数值、日期、布尔值等）的关系型数据表、数据库或数据仓库；**半结构化数据**主要是指有一定组织结构但包含自由格式数据的文件等，如网页集、论文集等等；**非结构化数据**则是指不能抽象出逻辑结构进行描述的数据文件，包括：纯文本、图形/图像（地图、遥感图像、空间数据）、音频(音乐、语音文件)、视频（影视、监控录像）、特殊应用数据（生物 DNA 序列）等。上述各种类型的数据所处的应用环境都可能会产生数据挖掘需求。基础的数据挖掘技术主要面向结构化的数据(包括用特征向量表示的纯文本、

有组织的 DNA 序列文件等）；而面向半结构化和非结构化的数据，则产生了实时/流数据挖掘、多媒体数据挖掘、社会网络挖掘、空间数据挖掘、图挖掘、多关系和多数据库挖掘、生物（DNA 序列）数据挖掘、语义挖掘等高级数据挖掘主题的研究与应用。

（2）应用原则

相对于其它算法，数据挖掘强调从真实的应用数据集中产生数据处理规则，再用这些规则来处理应用数据。这些规则的产生依赖于数据挖掘者对应用数据的业务含义有着清楚的认识。如果说对于统计类的算法，我们还可以抱着“先运算、再分析”的想法来看看会产生什么结果，那么对于数据挖掘类的算法，则必须要先想清楚要“挖”什么东西，策划好所用的“工具”（即哪种算法），再来动手实践。否则，难逃“garbage in，garbage out”(进去的是垃圾，出来的也是垃圾)命运。因而，数据挖掘应用要遵循以下几个原则：

1）充分的数据

数据的量一定要够，这是因为数据挖掘是以统计为基础，一定的样本基数是保证挖掘效果的前提。具体数据量的大小国算法而异，不同的算法对数据集大小的敏感程度不一样。

2）明确的数据

用于挖掘的数据一定要有明确的业务含义。对于二维表装载的数据，表中的每列数据的含义及列与列之间的关系要有清醒的认识。

3）合适的数据

数据格式要满足具体选用挖掘算法的格式要求。真实应用数据往往包含大量的噪声和冗余数据，原始数据集的存储格式、包含内容可能都不符合算法要求。因而对被挖掘数据对象的预处理工作往往占工作量的一半以上，具体内容包括噪声处理、数据归约、数据压缩、数据降维、泛化与规范化、格式转化等等。

4）清楚的目的

挖掘的目的一定要明确，确知要从数据中发现什么样的规律、规则，满足什么样的业务要求和目的。

5）适当的算法

在挖掘目的明确的情况下，首先要选择满足目的算法群，比如：寻找关联关系要选择关联分析类的算法，其次，要选择应用于数据集和运算环境的算法。这样才能起到事半功倍的效果。

4.4 支持电能质量高级分析

电能质量问题涉及广泛，引发电能质量问题的因素有电力设备、人为操作、自然灾害、电力网络布局、用户负载等多种，而电能质量引发的问题也有设备损坏、产能下降、安全事故、电价波动等等。

当前数据挖掘技术可以完成预测型数据分析任务；

当前数据挖掘技术可以在大量的多实体产生的数据中发现原来不为人所知的实体间的关联关系及影响关系，也能够自动地把相近的数据聚集在一起，以利于人们进行集中分析和利用；

当前数据挖掘技术对于电网中汇集的多节点数据可以进行汇总分析；

当前数据挖掘技术的发展潮流即是分布式计算，以应对大数据时代的到来；

当前数据挖掘技术有一支发展方向专门研究流数据，以满足实时处理数据的要求。

5 面向大数据：分布式数据挖掘

5.1 大数据与云计算平台

大数据(big data)，或称巨量资料，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯。大数据的 4V 特点:Volume、Velocity、Variety、Veracity。大的数据需要特殊的技术，以有效地处理大量的容忍经过时间内的数据。适用于大数据的技术，包括大规模并行处理（MPP）数据库，数据挖掘电网，分布式文件系统，分布式数据库，云计算平台，互联网，和可扩展的存储系统。

5.1.1 Hadoop 平台简介

Hadoop 平台是以分布式文件系统 HDFS(Hadoop Distributed Filesystem)和 MapReduce (Google MapReduce 的开源实现)为核心的，它为用户提供了系统底层细节透明的分布式计算和分布式存储的编程环境。Hadoop 简化了向集群中添加普通类型节点，并提供一套容错机制，其保留了数据的副本，具有高可靠性和可用性。因此，用户可以利用 Hadoop 轻松地组织计算机资源，进而搭建自己的分布式计算云平台，并且可以充分利用集群的计算和存储能力，完成海量数据的处理。

HDFS 的高容错性、高伸缩性等优点允许用户将 Hadoop 部署在低廉的硬件上，形成分布式文件系统；MapReduce 分布式编程模型允许用户在不了解分布式系统底层实现细节的情况下开发并行应用程序。

HDFS 将每个文件分割成默认大小为 64MB 的小块文件（大小可以根据用户需要来配置）。这些小块文件存储在集群的不同节点上。同时，为了保证可靠性，HDFS 将每个小块默认复制了 3 份。每块的元信息存储在 NameNode 名字节点的主存中。客户端从 NameNode 中查找块的位置然后到相关的 DataNode 数据节点中访问数据目录[7]。DataNode 要利用心跳检测机制频繁的向 NameNode 发送心跳，如果 NameNode 未能接收到来自 DataNode 的任何心跳，就将其检测为故障。

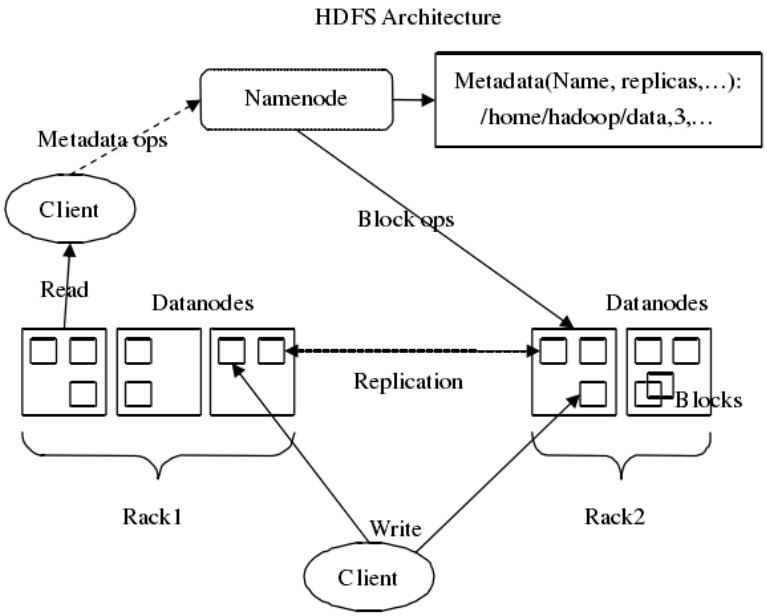


图 5 HDFS 总体架构图 [8]

MapReduce[9]是谷歌提出了一种并行化编程框架，其并行计算模式对任务的处理分为两个阶段：Map（映射）和 Reduce（规约），其对数据的处理流程如图 6 所示[9]。HDFS 中以 key-value 对形式存储的数据将作为 Map 阶段的输入。在 Map 阶段，根据指定的 InputFormat 类从文件块中读取数据并产生 key-value 对，然后调用自定义的 map 函数进行处理并产生中间结果存放在本地。在 Reduce 阶段，远程读取 Map 阶段产生的中间结果，调用自定义的 Reduce 函数进行处理，并将最终结果存储到 HDFS。用户只需根据具体情况实现 Map 和 Reduce 函数，这样极大的降低了用户并行化编程的难度。

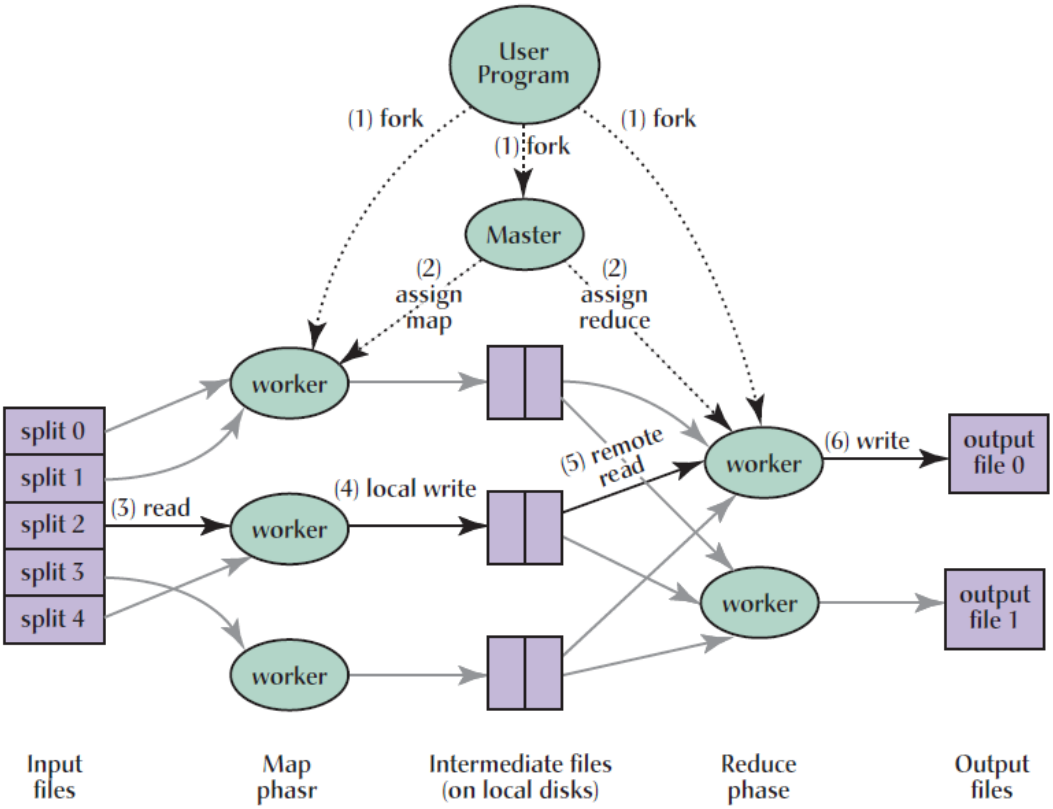


图 6 MapReduce 数据处理过程

5.2 分布式数据挖掘技术

大数据背景下，出现了很多分布式数据挖掘业务需求，例如对银行贷款的安全或者风险进行分类、多媒体数据关联分析、互联网舆情监控、推荐系统等。目前，大数据平台中主要采用 MapReduce 并行计算模型实现数据挖掘算法的并行化。该计算模型在对现有数据挖掘算法进行一定的改造后，更加适合于大数据分布式环境下数据挖掘业务，并且有助于异构数据的预处理和挖掘。

5.2.1 MapReduce 编程模式

Hadoop MapReduce 是为在平价的大规模集群上进行海量数据并行计算而设计的一款开源框架。MapReduce 编程模式解决了传统并行计算在易编程性方面存在的瓶颈。传统并行计算存在的问题，例如：工作调度、分布式存储、容错处理、网络通信等也都由 MapRcduce 负责解决。

（1）MapReduce 的体系结构

MapReduce 框架同 HDFS 一样采用 Master/Slave 架构，由主控节点(Jobtracker)和计算节点(Tasktracker)两种节组成。主控节点负责任务调度，为计算节点分配 Map 或 Reduce 任务，并且检测计算节点的执行愦况。如果某一任务执行失败，主控节点会重新分配节点执行该任务，以保证任务的完成。计算节点的职责是完成主控节点为其分配的任务。通常情况下，主控节点和管理节点可以使用同一台机器，计算节点和数据节点可以使用一群机器。这就意味着 HDFS 和 MapReduce 可以运行在同一个节点集之上。这种设计允许主控节点将任务分配给那些已经存储了数据块的计算节点(同时也是数据节点)。这样分配的优点是大节省了网络通讯的开销，提高了系统的效率。图 8 是将 HDFS 和 MapReduce 合并后的 Hadoop 架构图，其中 Master 节点负责管理节点和主控节点的工作，Slave 节点负责数据节点和计算节点的工作。

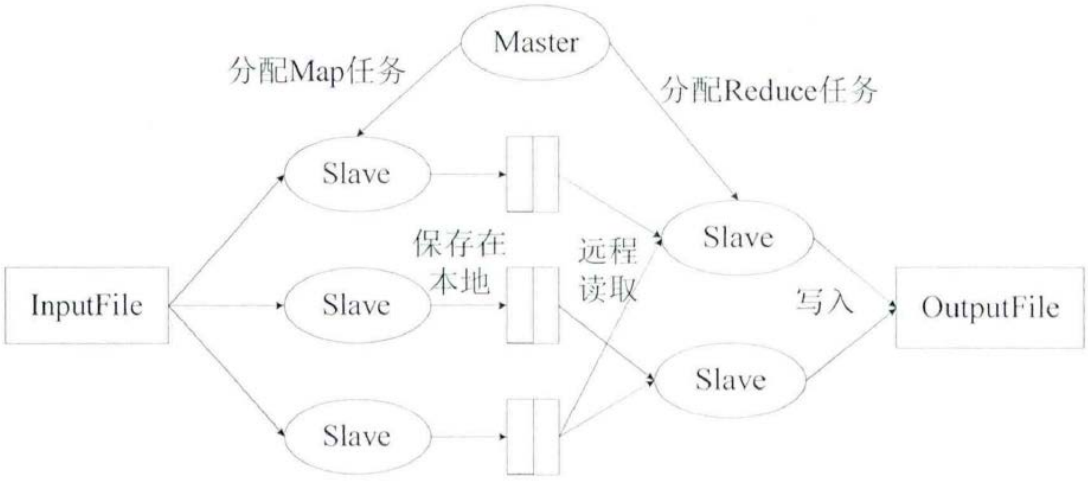


图 8 Hadoop 架构图

(2) MapReduce 的工作流程

MapReduce 的执行过程主要有以下几步:输入 Input 和切分 Split、映射 Map、洗牌 Shuffle、化简 Reduce、输出 Output。图 9 为 MapReduce 的工作流程图。

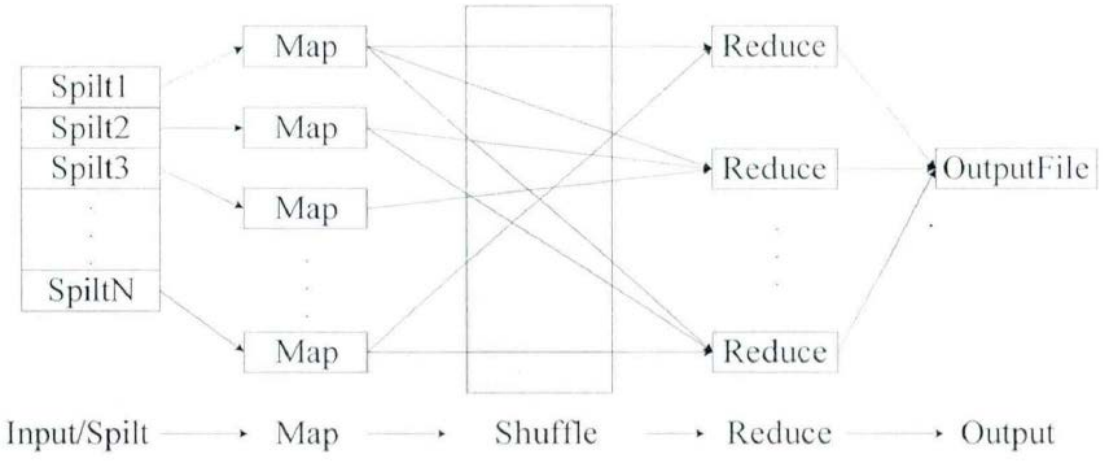


图 9 MapReduce 工作流程图

(1)输入 Input 和 切分 Split

在客户端的代码中要明确指出输入文件和输出文件的位置。在执行 Map 任务之前，输入文件会被划分为若干个 Split。每一个 Split 对应一个 Map 任务，该 Split 就是其对应的 Map 任务的输入。

注意:要区分这里的 Split 和 HDFS 中的 Block。Block 是物理上的划分，而 Split 是逻辑上的划分。在 MapReduce 启动前设置好 Split 的起始位置和切分长度后， MapReduce 会自动计算每个 Split 的位置。Split 初始默认值也是 64MB，也就是一个 Block 对应一个 Split。但是由于 Split 的长度是可配置的，所以同一个 Block 可能被划分到不同的 Split 当中，也有可能将多个 Block 划分到同一个 Split 当中。

(2)映射 Map

Map 任务将输入数据转换为中间结果输出，该中间结果会被保存在执行 Map 任务的本地节点上。该过程需要开发人员在容户端程序中编写 Map 函数，Map 函数的输入和输出数据格式都足键值对(<key，value>)集合。

MapReduce 允许输入键值对集合与输出键值对集合的类型、长度不同，一个给定的输入键值对可以映射到零个或多个类型不同的输出键值对。

(3)洗牌 Shuffle

Shuffle 是位于 Map 任务输出和 Reduce 任务输入之间的中间过程，它的作用是决定 Map 任务输出的每一个键值对会被发送到哪一个 Reduce 任务上作为输入。这一步实际上就是个哈希的过程，取 key 值的 Hashcode 对 Reduce 任务的数量取模(Reduce 任务的数量可以在 Hadoop 的配置文件中设定)。在这一步具有相同 key

值的输出键值对会被发送到同一个 Reduce 任务上。不同节点上的 Map 任务可能得到具有相同 key 的输出键值对，同一个节点上的 Map 任务也可能得到 key 不相同的输出键值对。

(4)化简 Reduce

Reduce 任务的输入和输出也都是键值对集合。Reduce 任务会接收--个 key 值和所有关联到这个 key 值的 value 組成的列表作为输入的键值对。通过编写客户端 Reduce 函数，对输入的键值对集合进行处理得到输出键值对集合。

(5)输出 Output

最后将多个 Reduce 任务的结果作为输出文件，输出到指泣的目录。

在上而的过程中省略了一个叫做“合并 Combine”的步骤，因为这个步骤不是必须的过程，而是可选的。Combine 任务的作用是优化 Map 任务生成的中间结果、减少 MapReduce 作业所使用的带宽。如果开发人员在客户端程序中编写了 Combine 函数，那么每一个执行 Map 任务的节点，都会在执行完 Map 任务后执行相应的 Combine 函数。然后将 Combine 函数的结果 Shuffle 到相应的 Reduce 任务上。

5.2.2 基于 Mahout 的分布式数据挖掘

Apache Mahout 是 ApacheSoftware Foundation (ASF) 旗下的一个开源项目，它是基于一个 Hadoop 的分布式计算框架，提供一些可扩展的机器学习领域经典算法的实现，可以辅助开发人员更好更快的开发数据分析程序。Mahout 基于 MapReduce 实现了一些数据挖掘算法，解决了部分并行挖掘的问题。Mahout 提供了一套具有可扩充能力的类库，它提供分布式计算框架的同时，还实现了一些可扩展的数据挖掘和机器学习领域经典算法，可以帮助开发人员更加方便快捷地创建智能应用程序。Mahout 中的算法能够高效地运行在分布式计算环境中，通过和 Apache Hadoop 分布式框架相结合，Mahout 可以有效地运行分布式数据挖掘算法[10]。

在 Mahout 实现的机器学习算法见下表：

算法类	算法名	中文名
分类算法	Logistic Regression	逻辑回归
	Bayesian	贝叶斯
	SVM	支持向量机

	Perceptron	感知器算法
	Neural Network	神经网络
	Random Forests	随机森林
	Restricted Boltzmann Machines	有限波尔兹曼机
聚类算法	Canopy Clustering	Canopy 聚类
	K-means Clustering	K 均值算法
	Fuzzy K-means	模糊 K 均值
	Expectation Maximization	EM 聚类（期望最大化聚类）
	Mean Shift Clustering	均值漂移聚类
	Hierarchical Clustering	层次聚类
	Dirichlet Process Clustering	狄里克雷过程聚类
	Latent Dirichlet Allocation	LDA 聚类
	Spectral Clustering	谱聚类
关联规则挖掘	Parallel FP-Growth Algorithm	并行 FP-Growth 算法
回归	Locally Weighted Linear Regression	局部加权线性回归
降维/维约简	Singular Value Decomposition	奇异值分解
	Principal Components Analysis	主成分分析
	Independent Component Analysis	独立成分分析
	Gaussian Discriminative Analysis	高斯判别分析
进化算法	并行化了 Watchmaker 框架	
推荐/协同过滤	Non-distributed recommenders	Taste(UserCF， ItemCF，SlopeOne)
	Distributed Recommenders	ItemCF
向量相似度计算	RowSimilarityJob	计算列间相似度
	VectorDistanceJob	计算向量间距离
非 Map-Reduce 算法	Hidden Markov Models	隐马尔科夫模型
集合方法扩展	Collections	扩展了 java 的 Collections 类

Mahout 最大的优点就是基于 hadoop 实现，把很多以前运行于单机上的算法，转化为了 MapReduce 模式，这样大大提升了算法可处理的数据量和处理性能。

推荐系统：根据用户过去的行为模式，尝试找出用户喜欢的事物 (例如购物或在线内容推荐)。

聚类分析：将对象按照相似性进行分组 （例如，找出类似专题的文档）

分类：学习现有类别的每个类别成员所具有的共同之处，在此基础上的分类尝试对新的对象进行分类。

繁项集挖掘，它建立一个对象组来标识通常在一起出现的对象。

6 面向流数据：实时流计算框架

6.1 流数据与实时计算

大数据时代，海量数据实时地、互不兼容地产生于社交网络、网络视频、传感器网络、电子商务等领域，很多新型应用需要长期地、持续地对数据进行流式处理、连续计算和分布式流数据处理等。“流数据”通常是指连续的、不间断的、非结构化的数据队列。因此，流数据可被视为一个随时间延续而无限增长的动态数据集合，通常具有以下特点：数据实时到达，输入顺序无法保证，数据的处理是一次性的，不是静态的存储后处理，而是动态的随到随时处理。数据经过处理后，如果不特意进行持久化操作，一般直接丢弃。

流计算是指一种高效地利用并行和定位，使用流计算处理器，流计算编程语言等多种技术手段处理流数据的新型计算模式[15,16]。不同于大数据中的面向非实时数据的批处理计算框架（例如 Hadoop），流计算面向的数据规模庞大且实时持续不断地到达，数据次序独立且时效性强，同时流数据的价值会随着时间的流逝而降低，要求数据在产生后必须立即对其进行处理。面对这种“大数据流”，传统的分布式计算模型不再能满足需求，而批处理计算框架在实时性、容错性等方面都有所欠缺。能够实时处理流动数据并作出合适决策的流计算技术应具备以下特点：

(1) 为了使得对数据流的处理延迟尽可能的低，系统必须摒弃将数据存储下来处理的传统做法，实时对数据进行处理然后丢弃。

(2) 流计算系统无法对数据流的产生和到达做任何假设，而且需要兼容静态数据和流数据，应用需要将当前数据与过去数据对比。

(3) 流计算需要具备在分布式节点上拓展的能力，各处理单元对实时数据进行提取、过滤和分析等操作和处理，单元只负责完成自身处理功能，每个处理单元地位平等，不能相互干涉。

(4) 流计算模型多核处理器和多线程的应用，充分发掘系统的计算资源，当负载增大时应当采用均衡技术使得负载转移到相对空闲的节点上去。

(5) 流计算能够实时整合来自多种异构数据源的数据，并对这些数据进行连续实时的处理。

目前，流计算的模型和框架成为业界研究的焦点，并已经形成了一系列分布式流计算框架，如 Yahoo! S4[17]，Facebook Data Free-way and Puma[18]，Twitter Storm[19]等等。

6.2 Simple Scalable Streaming System

Simple Scalable Streaming System（S4）是一个通用的、分布式、可伸缩的、部分容错的、可插拔的平台，该平台上开发人员可以很容易地开发应用程序来处理持续到达的流数据。S4 提供了简单的编程接口来处理流式数据，可以在普通硬件之上部署可扩展的高可用处理集群。S4 处理集群用于处理客户端发送的海量数据流量处理请求，集群包含很多分布式流计算处理节点，各节点之间完全对等，均具有相同的处理能力，同时每个处理节点内均含用户自定义的处理模块。，这种集群架构简化了部署和维护，提高了系统的可靠性和健壮性。同时，在 S4 框架中使用了可插拔的架构，各个模块之间松耦合，各模块即通用又可定制化。

S4 的分布式流计算处理集群采取流水线处理方式，各节点中不同的处理模块分别承担不同环节的处理工作，每个处理节点均各自独立运行，没有中心节点。各处理节点处理请求完成后，会将产生的中间结果发往其它的分布式节点继续进行处理，直到产生最终的处理结果，并将结构发送至客户端。

在 S4 中，所有发往分布式流计算处理集群的数据流会被封装成不同类型和属性值的事件，供分布式流处理集群进行处理。处理集群中，最基本的逻辑计算单元被称为一个 PEs (Processing Elements，处理单元)，各处理单元之间通过事件(events)进行交互。在一个 PE 收到或消费一个事件后，执行以下操作中的一项或两项：(1)发出一个或多个 event 给其它 PE。(2)公布结果。每个 PE 只消费事件类型、属性 key、属性 value 都匹配的事件(events)，并可能会产生输出事件。分布式流计算框架会为每个属性值初始化一个 PE，PE 负责监听某些特定类型的事件，可以通过指定更详细的“键-值”对来注册监听某类更细粒度的符合某种“键-值”关系的事件的特定子集。

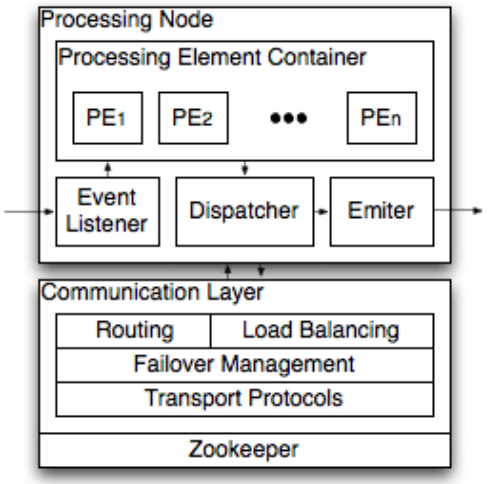
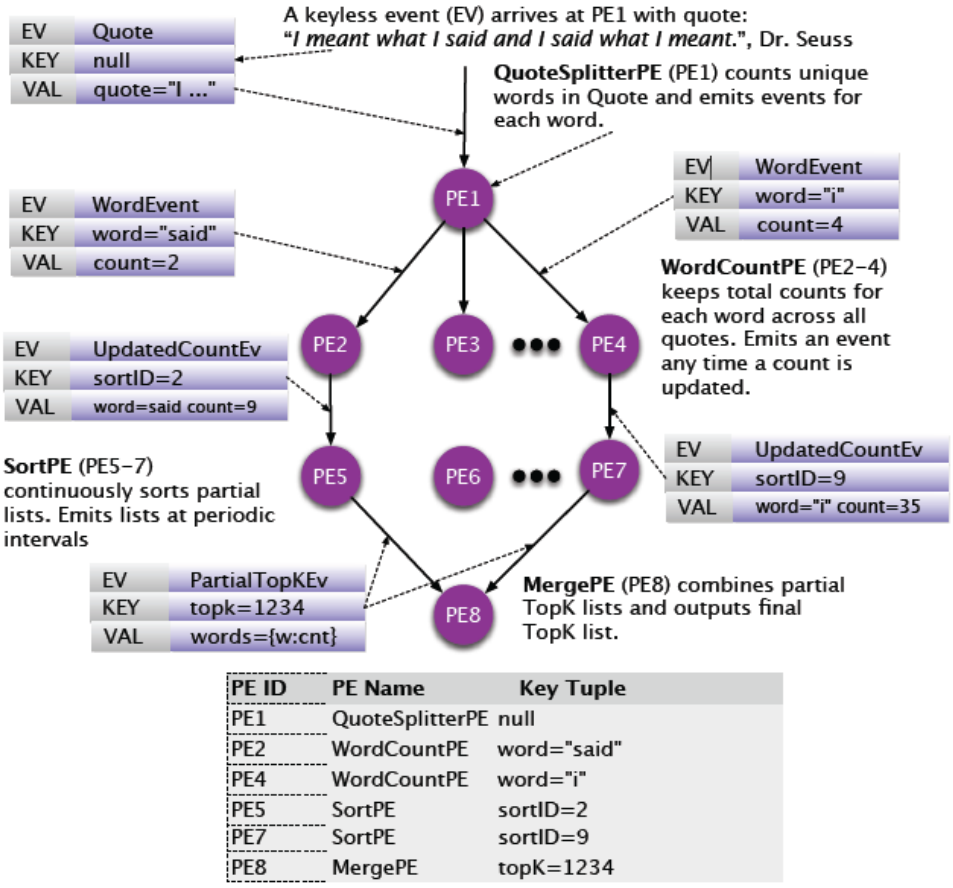


图 10 处理节点 Processing Node

处理节点 PN(Processing Node)是处理单元 PE 的逻辑主机，一个物理节点上可以有多个处理节点。如图 10 中所示，处理节点构建在通讯层的基础上，通信层提供了集群管理和容错功能，实现物理节点和逻辑节点的映射，通讯层能够自动检测到硬件故障和相应更新映射。在此基础上，处理节点负责监听事件，在到达事件上执行操作，通过通讯层的协助分发事件，并发送输出事件。S4 通过一个哈希函数将每个事件路由到处理节点，这个哈希函数作用于事件的所有已知属性值上。单个事件可能被路遇到多个 PN 上。所有可能的属性 Key 的集合通过 S4 集群的配置文件获知。PN 中的事件监听器将到来的事件传递给 PE 容器，PE 容器以适当的顺序调用适当的 PE。

在图 11 描述的示例中[20]，输入事件是一系列英文文档，任务是以最小的延迟持续的计算得出所有文档中出现最频繁的 K 个词。*Quote* 事件发送到 s4，该事件是没有 key 值的；然后，*QuoteSplitterPE* 对象(如图 2 中的 PE1)侦听系统中的 *Quote* 事件，*QuoteSplitterPE* 是一个无 key 值的 PE 对象，处理所有 *Quote* 事件。对于文档中每个单独的单词，*QuoteSplitterPE* 对象分配一个数，并发出新类型的事件 *WordEvent*，以 *word* 为 key 值。*WordCountPE*(如图 2 中的 PE2-4)以 *word* 为 key 值监听 *WordEvent* 事件。每次发现 key 值和 *word* 相匹配的单词，如果 *WordCountPE* 对象存在，PE 对象和计数器将递增，否则新 *WordCountPE* 对象实例化。当 *WordCountPE* 对象的计数器发生变化，它会将更新计数 *SortPE* 对象(如图 11 中的 PE5-7)。*SortPE* 对象的 key 值是一个[1, n]的随机整数，其中 n 是 *SortPE* 对象所需的数量。一旦 *WordCountPE* 对象选择一个 *sortID*，那么在它的整个生

命周期都会使用该 *sortID* 为发出事件的 Key 值。使用多个 *SortPE* 对象的目的是为了更好地将负载分布在多个处理节点或处理器。当 *UpdatedCountEvent* 事件到达时，每个 *SortPE* 对象更新 topK 列表，并定期地将其所属的 topK 列表发送 *PartialTopKEv* 消息到 *MergePE* 对象(如图 11 中的 PE 8)。*MergePE* 对象监听所有合并所有 *SortPE* 发出的 key 值为 topK 列表的事件，并合并各 topK 列表输出最终的全局 topK 列表。



PE ID	PE Name	Key Tuple
PE1	QuoteSplitterPE	null
PE2	WordCountPE	word="said"
PE4	WordCountPE	word="i"
PE5	SortPE	sortID=2
PE7	SortPE	sortID=9
PE8	MergePE	topK=1234

图 11 S4 单词计数流计算示例

6.3 Twitter Storm

Storm 与其他大数据解决方案的不同之处在于它的处理方式。Hadoop 在本质上是一个批处理系统。数据被引入 Hadoop 文件系统 (HDFS) 并分发到各个节点进行处理。当处理完成时，结果数据返回到 HDFS 供始发者使用。Storm 支持创建拓扑结构来转换没有终点的数据流。不同于 Hadoop 作业，这些转换从不停止，它们会持续处理到达的数据。

Storm 是一个分布式的、容错的实时计算系统，遵循 Eclipse Public License，可以方便地在一个计算机集群中编写与扩展复杂的实时计算。Storm 为分布式实

时计算提供了一组通用原语，可被用于流计算之中，实时处理消息并更新数据库。Storm 也可被用于“连续计算”（continuous computation），对数据流做连续查询，在计算时就将结果以流的形式输出给用户。Storm 还可被用于“分布式 RPC”，以并行的方式运行昂贵的运算。在流计算处理过程中，Storm 的主要特点如下：

a) Storm 使用 ZeroMQ 作为其底层消息队列，消除了中间的排队过程，使得消息能够直接在任务自身之间流动。

b) 在 Storm 框架之上可以使用多种常见编程语言，如 Clojure、Java、Ruby 和 Python 等，并且可以增加对其他语言的支持。

c) Storm 可以自动管理工作进程和节点的故障，具有很强的容错性。Storm 实现了有保障的消息处理，保证每个消息至少能得到一次完整处理，如果发现一个消息还未处理，会实现消息源重试消息。Storm 还实现了任务级的故障检测，在一个任务发生故障时，消息会自动重新分配以快速重新开始处理。

d) Storm 框架下，流计算可以在多个线程、进程和服务器之间并行进行的。

e) Storm 有一个“本地模式”，可以在处理过程中完全模拟 Storm 集群。这让你可以快速进行开发和单元测试。

f) Storm 集群由一个主节点和多个工作节点组成。主节点运行守护进程 Nimbus，工作节点运行守护进程 Supervisor。其中，Nimbus 的作用包括布置任务、分配代码及故障检测，Supervisor 的作用是监听和控制工作进程的启动与销毁。Nimbus 和 Supervisor 都具有很好的健壮性，它们都能够快速的从失败中恢复，而且是无状态的，它们的协调工作是由 Zookeeper 来完成的。

流计算中，Storm 框架中的角色通常包括 Stream、Tuple、Spout、Bolt、Topology、Stream Grouping 等，如图 3 所示。其中，Stream 是指被处理的流数据，这是一个无限的元组 Tuple 组成的序列。Storm 认为每个 stream 都有一个数据源 Spout，数据从外部来源 Spout 流入 Storm 拓扑结构中。Bolt 可以理解为流数据的中间状态转换，Stream Grouping 规定了 Bolt 可接收输入数据的类型，Topology 是由 Stream Grouping 连接起来的 Spout 和 Bolt 节点网络。

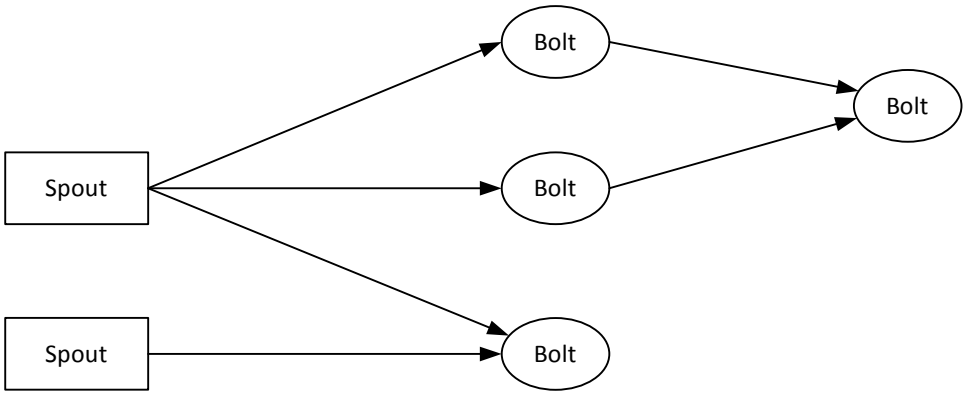


图 12 Storm 拓扑结构的概念架构

在图 12 描述的示例中[21]，*TweetsTransactionalSpout* 会连接到你的 tweets 数据库并且在 Topology 节点网络中作为数据源发射批量的元组。Topology 节点网络中包括两个不同的 Bolts，*UserSplitterBolt* 和 *HashtagSplitterBolt*，它们从 *TweetsTransactionalSpout* 接收元组。*UserSplitterBolt* 会分析 tweet 数据元组，查找用户名(以@为开头的词)，并发送这些词到一个叫做 *users* 的自定义流。*HashatagSplitterBolt* 同样分析 tweet 数据元组，查找标签词（以#为开头的词），并且发送这些词到一个叫做 *hashtags* 的自定义流。Topology 节点网络中还包括第三个 Bolt：*UserHashtagJoinBolt*，它接收 *users* 和 *hashtags* 两个流并且计算在一个命名用户的 tweet 中一个 hashtag 出现了多少次。最后，Topology 中还包括一个 Bolt 叫做 *RedisCommitterBolt*，接收由 *UserSplitterBolt*、*HashtagSplitterBolt* 和 *UserHashtagJoinBolt* 产生的流。在该示例中，它会完成所有计数，并且一旦完成了元组批次的处理就会发送到 Redis 数据库中。

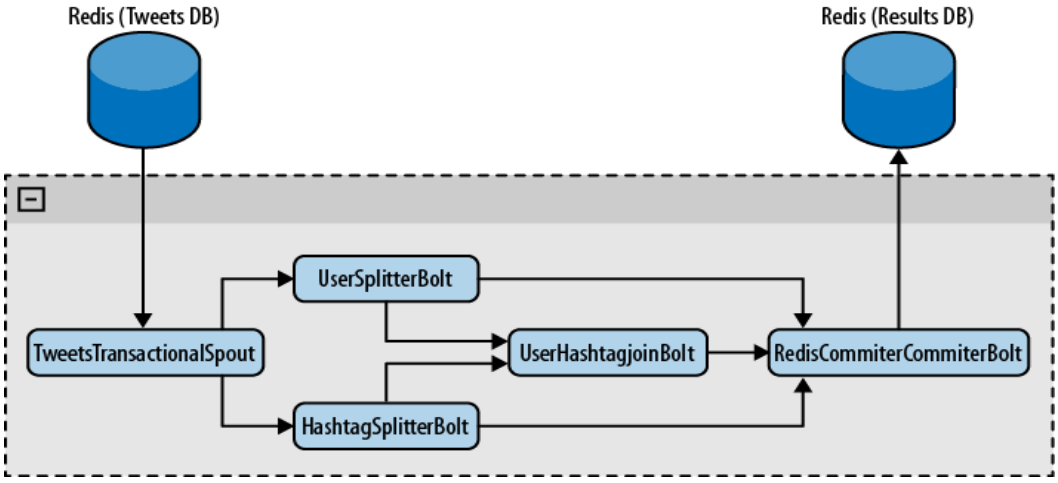


图 13 图 4 Topology 节点网络 [21]

7 面向预测：统计回归与时间序列分析

7.1 数据预测分析方法综述

预测是通过研究实物发展的历史和现状,运用科学的分析方法,综合各方面信息来揭示事物发展的客观规律,同时对事物的各种客观现象之间的作用做出科学分析,然后利用这些信息推测事物未来发展的可能途径和结果。预测方法很多,大体说来,可以分为技术方法和定量方法两大类。技术方法也称为定性方法,包括专家评估法、直观判断法等。定量方法主要利用原始数据,借助数学模型进行预测。常用的定量预测方法有回归分析预测、时间序列预测、趋势外推预测和灰色系统预测等目前,应用最广泛的预测技术有如下几种:

(1)回归分析预测

回归分析主要研究的是客观事物变量间的统计关系。依照不同的划分准则,同归分析有着不同的划分方法,按照自变量的个数可以将其分为一元回归和多元回归;按照回归方程的类型可将其分为线性回归和非线性回归;按照参数估计方法可以分为偏最小二乘回归、岭回归和主成分回归。

回归分析的一般方程为：$y=f(x_1,x_2,...,x_k)+\varepsilon$

其中 y 为因变量,x_1, x_2,…,x_k为自变量,$f(x_i,x_2,\cdots,x_k)$为变量 $x_1,x_2,\cdots,x_k$ 的确定性关系, ε为随机误差。

多元回归分析因其实用性和有效性,得到了广泛的应用。早在 1986 年,刊物《冶金地质动态》就发表了郑钟光[22]的多元回归分析在矿石体重测定中的应用,实践证明,通过建立回归方程来测定矿石的重量具有一定的优越性,可以提高工作效率和经济效益。张保国等[23]通过探讨影响老年人血压的社会学因素,对可能影响山东宵 60—69 岁老年人群的社会学指标进行逐步回归分析,从而获得影响血压的主导因素。刘伟铭等[24]使用多元回归分析法对高速公路事件持续时间进行研究,通过逐步回归分析得到了高速公路事件持续时间预测的最佳变量组合,同时建立了多元回归分析模型,经过验证得到预测结果与真实的事件持续事件情况基本相符。徐海量等[25]使用多元回归方法对引起塔里木河下游沙漠化的一些主要环境因子进行了分析,并得到影响塔里木下游沙漠化的两个重要限制因子,即地下水位和土壤含水量。

(2)时间序列预测

时间序列分析是数理统计的分支之一,根据动态数据揭示系统动态结构和规律,其基础是概率论与数理统计,支撑是计算机技术。任何时间序列经过合理的函数变换后都可以变为三部分的叠加,即趋势项部分、周期项部分和随机噪声部分。时间序列分析也就是从时间序列中把这三部分分解出來。时间序列分析的主要内容是依据系统有限长度的观测数据,建立能够比较精确反映时间序列中所包含动态依存关系的数学模型,然后利用模型的统计特性进行数据统计规律的分析,以达到预测或控制的目的。

时间序列分析的模型主要有自回归(AR)、滑动平均(MA)、自回归滑动平均(ARMA)、自回归滑动求和平均(ARIMA)模型。模型建立的基本方法是：首先利用时序图和相关系数來检验时间序列稳定性,倘若原始时间序列是不稳定的,则通过对数运算和差分运算将其转换为稳定的时间序列;然后对模型进行白噪声检验,只有满足非白噪声平稳的条件序列方能建立分析模型；再接着对模型进行识别、参数估计和模型检验,最后判断建立的模型是否合理。

近年来,时间序列已经成为一个相当活跃的领域。Daniel Billing[26]通过使用 ARIMA 模型对城市交通行程时间进行分析预测,实验表明,预测是有效的并且具有很好的可移植性,即可以方便地应用于其它地区的行程时间预测。Nayera Sadek等[27]使用分数阶ARIMA(F-ARIMA)模型对ATM的动态带宽分配使用情况进行预测,因为高速网络流量数据有着高度的自相关性,但是传统的AR和ARMA模型不能够获得长期的依赖特性,F-ARIMA 可以克服这一缺陷,能够通过数据获得短期和长期的依赖特性。结果表明,F-ARIMA 预测效果优于 ARIMA 模型。孙靖等[28]在空调负荷的预测研究中,通过季节差分消除了季节性周期性,再建立 ARIMA 模型对空调负荷进行预测,取得了较好的预测效果,预测结果可以很好地指导冰蓄冷系统的优化控制。张熙等[29]用模拟研究的方法,研究了如何填补念周期性的时间序列中随机型缺失数据,比较了周期性填补法和 spline 插值填补法,结果表明对于含有确定性周期性的时间序列,使用周期性填补法填补缺失数据的效果更好。

采用时间序列法进行预测,对于中短期预测效果较好,但是对于中长期预测有一定的局限性。

(3)趋势外推分析预测

趋势外推分析又称为趋势曲线分析、曲线拟合或曲线回归,即使用某条曲线来拟合获得的数据值,这条曲线能够反映数据的变化趋势,然后按照这个曲线的变化趋势,预测出未来某个时刻的数值。趋势外推函数常见的有:线性函数、抛物线函数、双曲线函数、指数函数、修正指数函数、幂函数等。

趋势外推分析预测的优点是仅仅依赖于过去的数掘值,所需的数据比较少。缺点是如果数据出现较大的波动,那么预测将出现较大的误差。使用趋势外推法进行预测分析的实例也很多,颜金木[30]利用趋势外推法对电力负荷进行了预测,但是这个应用是基于了两点假设,即负荷没有跳跃式变化和决定负荷发展的因素不变或者变化不大,虽然这样的分析预测有一定的可行性,但是也体现了趋势外推分析的缺点,负荷数据不能有突然的波动变化,否则预测结果会出现较大的误差。高志刚等[31]将趋势外推法应用于地表沉降预测中,取得了很好的效果,实例证明应用趋势外推法得到的预测模型能够很好地定量预测地表沉降的发展趋势。

(4)灰色系统预测

灰色系统理论是由中国学者邓聚龙教授在1982年创立的,灰色系统以部分信息明确、部分信息不明确的系统为研究对象,主要通过对明确信息进行生成、 开发、提取有价值的信息,从而实现对系统运行行为、演化规律的正确描述和有效监控。灰色系统的基本思想就是通过一定的处理将无明显规律的时间序列变成有规律的时间序列。灰色系统实质上是对原始序列进行累加生成,使生成的时间序列具有一定的规律性,再用一条光滑的曲线进行逼近

多年来,灰色系统预测在电力系统中长期负荷预测中的应用受到了广泛的关注,该方法利用灰色预测理论,结合电力系统负荷的特点,进行负荷预测。在九十年代以前,灰色系统预测主要局限于使用 GM(1,1)模型进行预测,但是预测效果不够理想;九十年代以后,陆续提出了改进方法,使得灰色系统预测法在应用中更为合理。刘思峰等[32,33]研究了 GM(1,1)模型的适用范围,通过发展系数阈值,明确界定了模型的有效区、慎用区、不宜区和禁区。姚天祥等[34]研究了离散GM(1,1)模型的特性,同时为解决灰色模型预测的病态性提供了思路,文章提出了分段修正离散GM(U)模型并对其机理进行了证明。

灰色系统预测法由于在建模时不需要计算统计特征量,因此理论上可以适用于任何非线性变化的负荷预测,但是也存在不足之处,其微分方程指数解比较适

合具有指数增长趋势的负荷预测,对于其它趋势的拟合灰度较大,精度的提高比较困难。

(5)专家系统预测

专家系统是一个基于知识的人工智能计算机程序系统,这些系统具有相当于某个专门领域的专家的经验和知识水平,以解决专门问题的能力。完整的专家系统是由知识库、推理机、知识获取部分和解释界面四部分组成的。

实践证明,精准的负荷预测仅仅依赖于高新技术和算法是不够的,还需要将人们长期积累的经验和大量的智慧融入进去。因此,专家系统这样的技术应运而生。但是目前专家系统的应用还不是非常广泛。阳春华等[35]利用专家系统对焦炉配煤进行了定性定量的设计,作者根据焦化理论和工业生产数据构造数学模型,同时以群体专家得到的定性知识构成规则模型,结合定量的数学模型和定性的规则模型,建立焦炭质量预测模型并实时控制配煤流量。此方法已经投入工业生产,并得到了很好的运行效果。

专家系统预测的应用还有一些实际的困难，比如：预测专家比较缺乏、预测过程还是容场出现人为的差错、在建数据库以及将专家经验转化为数学规则还存在系列的困难等等。这些问题还需要进一步探讨以找到解决方法。

(6)人工神经网络预测

“人工神经网络” (Artifical Neural Networks, ANN 或 NN)一词是相对于生物学中的生物神经网络系统而言的，其意图是使用一定的数学模型对生物神经网络系统进行描述再加之一定的算法，使其能在某种程度上模拟生物神经网络的智能行为，解决智能信息处理问题。人工神经网络的研究始于 1943 年,其发展经历了从兴起到萧条再到兴盛的曲折发展道路。通常，使用的人工神经网络模型有 BP 模型、Hopfield 模型、Kohonen 模型等,同时人工神经网络还可以与模糊集相结合构成模糊神经网络,用以对负荷预测中出现的模糊信息加以处理。

人工神经网络具有并行分布信息、非线性、自组织、自学习、自适应以及逼近任意连续函数的能力,这是常规算法和专家系统技术所不具备的。彭怀午等[36]使用人工神经网络对风电场短期功率预测进行了研究,结果表明,使用单一的 ANN 法预测精度还是比较低的,但将其与物理方法和统计方法相结合,预测精度得到了很大的提高。这也说明了人工神经网络存在一定的缺陷,其主要解决平稳

随机过程的预测,对于非平稳的过程预测效果不是很好,同时神经网络预测也存在学习速度慢,存在局部极小等固有缺陷。

(7)小波分析预测

小波分析是20世纪数学研究成果中最杰出的代表。小波分析是一种**时域——频域**分析方法，在时域和频域上同时具有良好的局部化性质。将小波理论运用到负荷预测中主要有两种思路：一是将负荷信号小波分解后根据各自的特性分别进行预测，再将预测到的信号进行重构，从而达到提高预测精度的目的。二是将其与神经网络理论相结合，建立小波神经元预测模型，通过小波神经元网络进行预测。

(8)组合模型预测

为了充分发挥不同预测方法的有用信息，提出了组合模型预测方法，从而尽可能地提高预测精度。组合模型预测就是先利用两种或两种以上的不同的预测法对同一预测对象进行预测,然后根据单独预测的结果选取适当的权值进行加权平均,最后取加权平均结果作为最终的预测结果。

郑鹏辉[37]在其硕士论文中提出了基于ARIMA模型的组合模型的研究,对时间序列模型、灰色预测模型以及BP神经网络模型进行了分析比较,在最小误差平方和的准则下,使用ARIMA模型和灰色模型的组合模型对中国GDP进行了预测分析,实验结果表明组合模型的预测精度优于每个单独预测的精度。

7.2 粗糙集理论引入预测分析

粗糙集理论(Rough Sets Theory)是1982年由波学者Pawlak提出,其是研究不确定、不完整的知识和数掘表达、归纳和学习的数学理论方法[38]。最初的时候由于受到语言限制,只有东欧等国的部分学者研究,之后才慢慢地被国际上的数学界和计算机界重视。1991年,Pawlak在其出版的专著《粗糙集---关于数据推理的理论》[39]中首次系统地阐述了粗糙集理论,从而将粗糙集理论及其应用的研究带入了一个新的阶段。1992年Slowinski R主编的论文集《智能决策支持:粗糙集理论应用与发展手册》出版[40]，同年在波兰Kiekrz召开第一届国际粗糙集研讨会,从此每年都会召开一次以粗糙集理论为主题的国际研讨会,从而推动了粗糙集理论的拓展和应用。

国内对粗糙集理论的研究相对晚些,国内最早的 RS 专著《粗糙集理论及其应用》由曾黄霖教授在 1998 年编著。目前国内有关粗糙集理论的著作主要还有王国澈的《Rough 集理论与知识获取》,刘情的《Rough 集与 Rough 推理》,苗夺谦等。这些著作对我国粗糙集理论的研究与发展起到了巨大的推动作用。

粗糙集理论的重要思想是将不确定的知识用已知的知识库来近似描述,其区别于其它处理不确定问题的理论主要是 RS 理论不需要除了所处理的数据集之外的任何先验信息,所以其对不确定问题的处理是比较客观的[41]。

知识约简是 RS 理论的核心内容之一，在智能信息或数据处理中占有十分重要的位置。通常,知识库中的知识并不是同等要的,有些甚至是冗余的。所谓的知识约简就是在保持知识库的分类能力不变的条件下,删除其中不必要的知识。知识约简包括属性约简和值约简(又称为决策规则的获取)[39]。

粗糙集理论的主要优点除客观性之外,其优势还在于:能够搜集数据的最小集合;评价数据的重要性;可以处理定性和定量数据;从数据中产生决策规则的集合[42]。粗糙集理论不仅可以用于构造新型系统,同时能优化现有的许多算法。粗糙集的决策推理规则能够用于预测分析,同时可将粗糙集和神经网络、回归分析、灰色理论等结合起來进行分析预测。

雷绍兰等[43]提出使用模糊粗糙集方法对电力负荷进行预测分析,首先结合模糊集和粗糙集的特性,提出一种小区划分和空间属性规则提取的方法,由于小区划分会导致模糊推理规则的成倍增加,因此采用粗糙集理论的属性约简方法约简预测因子同时得到决策推理规则库;使用粗糙集重要度分析计算不同条件属性的权重,克服模糊集确定权重的主观性。张宏刚等[44]将粗糙集理论与时间序列分析理论相结合,对电力负荷进行预测。文中首先使用粗糙集理论对影响负荷的气象因素进行约简,找到核心气象因素,然后利用时间序列分析方法进行预测,实践证明,这是一种适用性很强的技术。钟波等[45]将粗糙集和神经网络分析模型结合起来,建立了一种新型的预测模型。文中融合粗糙集方法与神经网络方法各自的优势,首先利用粗糙集约简了负荷的影响因素,小区冗余倍息、简化了网络输入变量,获得典型样本。然后用典型样本约简隐含层神经元和训练网络，并将网络连接权值学习的非线性问题转化为线性问题，是网络结构得到优化，数值试验说明该模型是可行、有效、实用的。费胜巍等[46]将粗糖集与灰色理论结合,用于电力变压器故障的预测。首先根据粗糙集的特点获得改进的三比值诊断决策表,

并通过简化决策表建立最小诊断规则；分别建立决策表的三比值的灰色预测模型，通过灰色模型对特征气体的比值进行预测，将获得的特征气体的状态特征与最小诊断规则相对照，得到预测的故障类型。

目前，将粗糙集理论引入回归分析的理论也很多。张诚等[47]利用粗糙集和多元回归分析对江西铁路物流需求进行预测，文中先用粗糙集的属性约简理论对自变量进行了筛选，然后利用筛选后的变量建立回归分析模型，但是作者没有给出针对多重共线性问题的解决方法，对回归模型的显著性检验也很模糊。罗仁吉[48]在其硕士论文单井投资估算与经济效益评价中，利用粗糙集对条件属性进行了约简，并生成决策规则，文中没有将粗糙集理论和回归分析融合起来使用。刘盾[49]将粗糙集属性约简理论引入了线性回归分析，作者主要提供了一个思路，先利用粗糙集理论对条件属性进行筛选，再建立回归分析方法，然后将建立的回归分析方程与利用逐步回归分析法得到的回归方程进行比较，也就是将逐步回归分析作为一种检验方法，但是这种方法有个缺陷，就是必须依赖于逐步回归分析。同时，在本文的实验中发现，利用 SAS 分析软件进行逐步回归分析时，显著性检验水平默认为 0.15，而常用的显著性水平为 0.05，这也给分析研究选用不同的显著性水平进行检验带来一定的不便。

夏丽[50]将粗糙集理论与回归分析理论结合起来进行横向分析，既充分发挥了粗糙集属性约简的优势，属性约简可以消除冗余属性、剔除相对不重要的属性，同时得到的约简属性能够保持整个信息系统的结构不变。同时利用回归分析的显著性检验来分析属性约简得到的属性之间是否存在多重共线性。在同时满足显著性检验和约简属性条件时，也就是选出了的自变量满足个数最少且能表达全部的信息系统同时解决了多重共线性问题。同时，将粗糙集引入回归分析还有一个原因是粗糙集对定量数据和定性数据都可以很好地分析，因为该文是通过建立回归模型进行预测分析，所以这一优势并没有能够显示出来。对于自变量是定性值的利用粗糙集也能进行筛选，选出对因变量影响相对比较大的自变量。粗糙集对自变量的筛选的应用范围更广，因此将粗糙集引入回归分析值得进一步研究。

7.3 回归分析预测

客观世界中的许多关系,例如:人的身高和体重之间的关系、人的血压和年龄之间的关系等,它们之间是相关的,但是其关系又不能用普通的函数来表示。这种具有相关关系的变量之间虽然没有确定的函数关系,但是可以借助函数关系表达它们之间的统计规律,这种近似地表达相关关系的函数被称为回归函数。

7.3.1 多元线性回归分析

回归分析分为线性回归分析和非线性回归分析,线性回归分析又分为一元线性回归分析和多元线性回归分祈。线性回归分析在整个回归分析中是最重要的,主要有两个原因:一方面是在客观世界中,线性回归分析的应用是最广泛的;另一方面是只有在回归分析为线性的假设条件下,才能得到比较深入和一般的结果,并且许多非线性回归问题可以转换为线性问归分析问题。因此,线性回归分析是回归分析研究的重点。一元线性回归是线性回归的一种特例,下面着重介绍多元线性回归分析。

多元线性回归是研究两个或两个以上的自变量与一个因变量的是否存在相互依存关系(线性关系)。通常可以用多元回归方程刻来表达这种关系,即多元方程刻画的是一个因变量与多个自变量之间的关系[51]。方程式中有两个或两个以上自变量的线性回归模型称为多元线性回归模型。在该模型中,因变量 Y 是多个自变量 $X_1, X_2, \cdots, X_k$和误差项的线性函数,表达式如下[52]:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k \tag{7-1}$$

对随机误差项ε常假定 E(ε)=0,Var(ε)=σ^2。并且称

$$E(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k \tag{7-2}$$

为理论回归方程。

在实际应用中,如果获得 n 组观测数据($X_{i1}, X_{i2}, \cdots, X_{ik}; Y_j$),i=1,2,…,n,则线性回归模型变为:

$$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_kX_{ik}+\varepsilon_i \tag{7-3}$$

其相应的矩阵表达式为 $Y=X\beta+\varepsilon$ (7-4)

其中

$$Y=\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \quad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \varepsilon=\begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

为了便于模型的参数估计,对方程(7.3)式有如下一些基本假定:

(1)解释变量 $X_1, X_2, \cdots, X_k$是确定性变量,X 是一满秩組阵;

(2)随便误差项具有零均值和等方差,即满足 Gauss-Markov 条件:

$$\begin{cases} E(\varepsilon_i)=0,i=1,2,\cdots,n \\ Cov(\varepsilon_i,\varepsilon_j)=\begin{cases} \sigma^2,i=j\ (i,j=1,2,\cdots,n) \\ 0,i\neq j \end{cases} \end{cases} \qquad (7\text{-}5)$$

(3)随机误差项服从正态分布:

$$\varepsilon_i \sim N(0,\sigma^2),i=1,2,\cdots,n \qquad (7\text{-}6)$$

对于多元线性回归的矩阵形式(7-4)式,这个条件为:

$$\varepsilon \sim N(0,\sigma^2 I_n) \qquad (7\text{-}7)$$

由以上假设和多元正态分布的性质知,Y 服从 n 维正态分布:

$$Y \sim N(X\beta,\sigma^2 I_n) \qquad (7\text{-}8)$$

7.3.2 多元非线性回归分析

在实际应用中,很少有变量与变量之间是成绝对的线性关系,线性回归分析并不适用于所有的情形,因此有必要了解非线性回归分析，从线性回归模型过渡到非线性回归模型，可以用非线性回归分析來建立变量之间的关系。常见的非线性函数有:①多项式 $y=a+bx+cx^2+\ldots+mx^n$；②双曲线 $\frac{1}{y}=a+\frac{b}{x}$ ③指数函数 $y=a\exp(bx)$ ④幂函数 $y=ax^b$ ⑤对数函数 $y=a+b\ln x$ 等[53]。

非线性回归分析中使用最多的是多项式回归分析,即研究一个因变量与一个或多个自变量间多项式的回归分析。例如:一元 m 次多项式回归方程为

$$\hat{y}=b_0+b_1x+b_2x^2+\cdots+b_mx^m \qquad (7\text{-}9)$$

可以通过逐渐增加X的高次项对实测点进行逼近,直至满意为止,这也是多项式回归的最大优点。在大量实际问题中,不论因变量与其它自变量的关系如何,总可以用多项式回归来分析。

在建立了多项式回归的方程后,可以通过变量的转换来进行求解。在(7-9)的方程中,可以令 $x_1=x, x_2=x^2, \cdots, x_n=x^n$,则可以将上述方程转化为m元线性回归方程，即

$$\hat{y} = b_0+b_1x_1+b_2x_2+\cdots+b_mx_m \qquad (7\text{-}10)$$

在多项式回归中需要检验回归系数 b_i 是否显著，实质上就是自变量x的i次方项 x^i 对因变量y的影响是否显著。

对于有多个变量的回归分析步骤如下：

(1)绘制散点阁,观察散点阁的分布特性；

(2)按照所选取的函数进行相应的变量转换；

(3)对变换后的数据建立线性回归模型；

(4)拟合多个相近模型,然后通过比较各个模型的拟合优度选择合适的模型。

7.4 时间序列分析

对时间序列进行分析,就足要根据这些时问序列,较精确地找到相应系统内在统计特性和发展规律性。目前对时问序列数据进行分析和预测比较完善和精确的算法是博克思一詹金斯(Box-Jenkins)方法,其基本思想为:时间序列是一组依赖于时间的随机变量，这组随机变量之间具有的依存关系或者相关特性表明了预测对象发展的延续性,将其中所蕴含的自相关特性使用数学模型描述出来,就可以利用时间序列的过去值和现在值来预测其将来的值。

7.4.1 时间序列的定义

时间序列是一个有序的观测值序列,通常是按照时间观测的,但也可以按照其它度量來观测,如长度、温度、速度等等,其中时间间隔可以是等间隔也可以是非等间隔[54]。

7.4.2 时间序列的分类

时间序列根据所研究的依据的不同,有着不同的分类[55][56][57]。

以所研究的对象的个数为依据,可分为一元时间序列和多元时间序列。研究的对象为一个变量的序列称为一元时间序列;研究的对象是多个变量的序列称为多元时间序列。研究多元时间序列不仅仅要分析单个变量随着时间的变化规律,同时要揭示各变量相互依存关系的动态规律性。

以时间的连续性为依据,可分为连续时间序列和离散时间序列。连续时间序列是指每个序列值所对应的时间参数为连续函数;离散时间函数是指每个序列值所对应的时间参数为间断点。对于连续数据,可以通过离散化算法使之转化为离散时序数据。

以序列的统计特性为依据,可分为平稳时间序列和非平稳序列。非平稳序列通常包含趋势性、季节性或周期性等的一种或者几种。对于非平稳序列在适当的时间去掉趋势性和季节性后,剩下的周期性部分通常会有某种平稳性。平稳序列通常分为严平稳序列和宽平稳序列。其中,严平稳序列是一种比较苟刻的平稳性定义,它认为时间序列的概率分布与时间 t 无关,那么这样的时间序列称为严平稳序列。若序列$\{Y_t\}$的一、二阶矩存在,同时在任意时刻 t 满足:

(1)$E\{Y_t\}=\mu$,μ为常数, $t\in T$;

(2)$E\{(Y_{t+k}-\mu)(Y_t-\mu)\}=r_k$, $t,t+k\in T$。

则称这样的时间序列为宽平稳时间序列,也叫做广义平稳时间序列。通常,研究的时间序列主要是宽平稳时间序列。

7.4.3 时间序列的数字特性

对于平稳的时间序列,采用四个变量对它进行描述,即均值μ、方差σ^2、自协方差γ^k和自相关函数ρ^k。其中均值$E(x_t)=\mu$,方差$Var(x_t)=E(x_t-\mu)^2=\sigma^2$,协方差$Cov(x_t, x_{t+k})=E[(x_t-\mu)(x_{t+k}-\mu)$,自相关函数为 $\rho_k=\gamma_k/\gamma_0$。

在多数实际应用中,不可能根据定义计算出均值、方差、自相关函数和偏相关函数,所以只能根据样本$\{Y_t\}$的时间平均来估计,也称为样本均值、样本方差、样本自协方差和样本自相关函数。它们的计算如下:

样本均值: $\overline{Y}=\frac{1}{n}\sum_{t=1}^{n}Y_t$

样本方差: $\hat{\sigma}_Y^2=\frac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})^2$

样本自协方差: $\hat{\gamma}_k=\frac{1}{n}\sum_{t=1}^{n-k}(Y_t-\overline{Y})(Y_{t+k}-\overline{Y})$

样本自相关函数: $\hat{\rho}_k=\frac{\hat{\gamma}_k}{\hat{\gamma}_0}$

7.4.4 时间序列分析建模步骤

运用Box-Jenkins分析方法进行预测的前提条件是预测对象是平稳的随机序列,因此在分析和建立合适的模型之前,必须对时间序列进行平稳化处理。预测的基本步骤为:

(1)利用自相关函数和偏自相关函数等方法进行分析,即分析时间序列的随机性、季节性和平稳性,并选择一个合适的模型来拟合所分析的时间序列数据,即模型识别阶段。

(2)通过时间序列数据对模型的参数进行估计,并对模型进行检验,判断建立的模型是否合适。如果不合适,须重新进行模型的识别,重新选择模型。

(3)对未来的数值进行预测,并分析预测的准确率。

7.4.5 时间序列分析发展趋势

近几年时间序列分析研究在理论上取得了很大的进展,同时被应用到不同领域并取得了令人满意的分析效果。但是受到某些方面的局限,时间序列分析模型和时间序列分析方法还有值得改进和继续发展之处,也是时间序列分析发展的未来趋势。时间序列分析的未来发展趋势主要集中于以下几个方面[58][59]:

(1)多变量时间序列。在不同的科学领域,多变量时如序列(MTS)集是很常见的,因此对其的分析和研究有着很广泛应用领域。对于多变量时间序列可以通过分析不同变量间的相关性來提取更多有价值的信息并充分分析这些信息以取得更好的预测效果。因此对多变量时间序列数掘进行分析和建模有着重要的价值。

(2)神经网络分析。目前有大量的分析预测技术,但由于时间序列信息的不完整性和各种因素的影响,那么建立一个有着智能信息处理能力的预测系统是很有必要的,将神经网络技术运用到时间序列中将会是一个发展趋势。同时为获得更加精确的预测结果将模糊集、遗传算法与神经网络算法进行整合也将是发展趋势。

(3)数据预处理理论。众所周知,随着信息时代的到来,在各行各业每天都面对着大量的数据。但是,这些数据也包含着各式的错误,比如冗余数据、数据丢失、

不确定数据、无关紧要的数据等等。这些都阻碍这数据的分析甚至会影响预测结果的准确性。因此,为了提高数据挖掘的准确性和降低数据处理的数量,在对数据进行挖掘之前的数据预处理就显得至关重要了。所以,如何有效地对海量数据进行预处理在未来的研究中有着重要的位置,这也是对时间序列分析研究的趋势之一。

(4)时间间距问题。目前,研究接触到时间序列数据多为等间距时间序列,那么对非等间距时间序列如何处理也将是未来研究的趋势之一。

最后,由于现实生活中大量的时间序列真实模型都是非平稳、非线性的,用线性方法处理问题是就存在不可避免的缺陷和局限。因此,对现实数据的研究中,找到具有良好的非线性品质、极高的拟合精度是很必要的,这样才能从现实应用出发,使得研究的时间序列具有更高的预测精度、更强的实际应用价值,达到更好解决实际问题的目的。

8 面向复杂决策：分类技术

8.1 分类、预测与决策

分类是一种重要的数据挖掘技术。分类的目的是根据数据集的特点构造一个分类函数或分类模型（也常常称作分类器），该模型能把未知类别的样本映射到给定类别中的某一个。构造模型的过程一般分为训练和测试两个阶段。在构造模型之前，要求将数据集随机地分为训练数据集和测试数据集。在训练阶段，使用训练数据集，通过分析由属性 描述的数据库元组来构造模型，假定每个元组属于一个预定义的类，由一个称作类标号属性的属性来确定。训练数据集中的单个元组也称作训练样本，一个具体样本的形式可为：（$u_1, u_2, \cdots\cdots u_n$;c）；其中 u_i 表示属性值，c 表示类别。由于提供了每个训练样本的类标号，该阶段也称为有指导的学习。通常，模型用分类规则、判定树或数学公式的形式提供。在测试阶段，使用测试数据集来评估模型的分类准确率，如果认为模型的准确率可以接受，就可以用该模型对其它数据元组进行分类。一般来说，测试阶段的代价远远低于训练阶段。

分类和回归都可以用于预测。和回归方法不同的是，分类的输出是离散的类别值，而回归的输出是连续或有序值。分类也可用于决策支持，其中最典型的应用为基于分类决策树的决策支持方法。基于分类技术进行决策支持，可以在面向数据种类繁多、决策分支明确的情况下，实现对于大数据量生产环境的快速行为决策，并最大限度减少人为的主观因素影响。

8.2 数据预处理

为了提高分类的准确性、有效性和可伸缩性，在进行分类之前，通常要对数据进行预处理，包括：

(1) 数据清理：其目的是消除或减少数据噪声，处理空缺值。

(2) 相关性分析：由于数据集中的许多属性可能与分类任务不相关，若包含这些属性将减慢和可能误导学习过程。相关性分析的目的就是删除这些不相关或冗余的属性。

(3) 数据变换：数据可以概化到较高层概念。比如，连续值属性“收入”的数值可以概化为离散值：低，中，高。又比如，标称值属性“市”可概化到高

层概念“省”。此外,数据也可以规范化,规范化将给定属性的值按比例缩放,落入较小的区间,比如[0,1]等。

8.3 分类算法的种类及特性

分类算法的目标是将未知样本归类到不同的分组中，通过学习各种统计数据中类别和样本之间的关系创建一个模型（通常称为分类器），以便以后用于归类未标识的样本。从机器学习的概念来说，聚类算法是一种无监督的学习算法，它自动地决定样本被分至的类别，而分类算法则是一种监督学习算法，通过模仿和学习样本的正确分类来训练分类模型。

分类模型的构造方法有决策树、统计方法、机器学习方法、神经网络方法等。按大的方向分类主要有：决策树，关联规则，贝叶斯，神经网络，规则学习，k一临近法，遗传算法，粗糙集以及模糊逻辑技术。

8.3.1 决策树分类算法

决策树（Decision Tree）是一种有向无环图（Directed Acyclic Graphics, DAG）。决策树方法是利用信息论中的信息增益寻找数据库中具有最大信息量的属性字段,建立决策树的一个节点,再根据该属性字段的不同取值建立树的分支,在每个子分支子集中重复建立树的下层节点和分支的一个过程。

构造决策树的具体过程为：首先寻找初始分裂，整个训练集作为产生决策树的集合，训练集每个记录必须是已经分好类的，以决定哪个属性域（Field）作为目前最好的分类指标。一般的做法是穷尽所有的属性域，对每个属性域分裂的好坏做出量化，计算出最好的一个分裂。量化的标准是计算每个分裂的多样性（Diversity）指标。其次，重复第一步，直至每个叶节点内的记录都属于同一类且增长到一棵完整的树。

主要的决策树算法有 ID3、C4.5(C5.0)、CART、PUBLIC、SLIQ 和 SPRINT 算法等。它们在选择测试属性采用的技术、生成的决策树的结构、剪枝的方法以及时刻，能否处理大数据集等方面都有各自的不同之处。

（1）ID3 算法

在当前决策树学习的各种算法中，影响最大的是 J R.Quinlan 于 1986 年提出的 ID3 算法，他提出用信息增益作为属性的选择标准，以使得在对每一个非叶

节点进行测试时，能获得关于被测试记录最大的类别信息。ID3 总是选择具有最高信息增益的属性作为当前节点的测试属性。具体方法是：检测所有的属性，选择信息增益最大的属性产生决策树节点，由该属性的不同取值建立分支，再对各分支的子集递归调用该方法建立决策树节点的分支，直到所有子集仅包含同一类别的数据为止，最后得到一棵决策树，它可以用来对新的样本进行分类。ID3 算法通过不断的循环处理，初步求精决策树，直到找到一个完全正确的决策树。在选择重要特征时利用了信息增益的概念。

设 S 是 s 个数据样本的集合。假定类标号属性具有 m 个不同值，定义 m 个不同类 $C_i(i=1,\cdots,m)$。设 s_i是类 C_i中的样本数。对一个给定的样本分类所需的期望信息由下式给出：

$$I(s_1,s_2,\cdots,s_m)=-\sum_{i=1}^{m}p_i\log_2(p_i)$$

其中 $p_i=s_i/s$ 是任意样本属于 C_i的概率。注意，对数函数以 2 为底，其原因是信息用二进制编码。

设属性 A 具有 v 个不同值$\{a_1,a_2,\cdots\cdots,a_v\}$。可以用属性 A 将 S 划分为 v 个子集$\{S_1,S_2,\cdots\cdots,S_v\}$，其中 S_j中的样本在属性 A 上具有相同的值 $a_j(j=1,2,\cdots\cdots,v)$。设 s_{ij}是子集 S_j中类 C_i的样本数。由 A 划分成子集的熵或信息期望由下式给出：

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+\cdots+s_{mj}}{s}I(s_{1j}+\cdots+s_{mj})$$

熵值越小，子集划分的纯度越高。对于给定的子集 Sj，其信息期望为

$$I(s_{1j}+\cdots+s_{mj})=-\sum_{i=1}^{m}p_i\log_2(p_i)$$

其中 $p_{ij}=s_{ij}/|s_j|$是 S_j中样本属于 C_i的概率。在属性 A 上分枝将获得的信息增益是:

$$Gain(A)=I(s_1,s_2,\cdots,s_m)-E(A)$$

该算法优点在于：（1）算法的基础理论清晰，方法简单，计算速度快；（2）搜索空间是完全的假设空间，目标函数就在搜索空间中，不存在无解的危险；（3）全盘使用训练数据，可得到一棵较为优化的决策树。

在实际应用中，对于非增量式的学习任务，ID3 算法通常是建立决策树的很好选择，但该算法不足之处在于：（1）不能增量地接受训练例，这就使得每增加一次实例都必须废除原有的决策树，重新计算信息增益并构造新的决策树，这

造成极大的开销；（2）只能处理离散属性，在分类前需要对其进行离散化的处理；（3）在建树时，每个节点仅含一个特征，这是一种变元的算法，特征间的相关性强调不够；（4）对噪声较为敏感，数据质量差将直接导致生成的决策树过于庞大或决策树中很多分支的信息量很少。（5）在建树的过程中每当选择一个新属性时，算法只考虑了该属性带来的信息增益，未考虑到选择该属性后为后续属性带来的信息增益，即未考虑树的两层节点；（6）其信息增益存在一个内在偏置，它偏袒属性值数目较多的属性。

(2) C4.5 算法

C4.5 算法继承了 ID3 算法的优点，并在以下几方面对 ID3 算法进行了改进：

1) 用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足；

2) 在树构造过程中进行剪枝；

3) 能够完成对连续属性的离散化处理；

4) 能够对不完整数据进行处理。

C4.5 算法与其它分类算法如统计方法、神经网络等比较起来有如下优点：产生的分类规则易于理解，准确率较高。其缺点是：在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，因而导致算法的低效。此外，C4.5 只适合于能够驻留于内存的数据集，当训练集大得无法在内存容纳时程序无法运行。

（3）SLIQ 分类算法

针对 C4.5 改进算法而产生的样本集反复扫描和排序低效问题，SLIQ 分类算法运用了预排序和广度优先两项技术进行了改进。预排序技术消除了节点数据集排序，广度优先策略为决策树中每个叶子节点找到了最优分裂标准。

1) 预排序。对于连续属性在每个内部节点寻找其最优分裂标准时，都需要对训练集按照该属性的取值进行排序，而排序是很浪费时间的操作。为此，SLIQ 算法采用了预排序技术。所谓预排序，就是针对每个属性的取值，把所有的记录按照从小到大的顺序进行排序，以消除在决策树的每个节点对数据集进行的排序。具体实现时，需要为训练数据集的每个属性创建一个属性列表，为类别属性创建一个类别列表。

2) 广度优先策略。在 C4.5 算法中，树的构造是按照深度优先策略完成的，需要对每个属性列表在每个节点处都进行一遍扫描，费时很多，为此，SLIQ 采

用广度优先策略构造决策树，即在决策树的每一层只需对每个属性列表扫描一次，就可以为当前决策树中每个叶子节点找到最优分裂标准。

SLIQ 算法由于采用了上述两种技术，使得该算法能够处理比 C4.5 大得多的训练集，在一定范围内具有良好的随记录个数和属性个数增长的可伸缩性。

然而它仍然存在如下缺点：

1)由于需要将类别列表存放于内存,而类别列表的元组数与训练集的元组数是相同的,这就一定程度上限制了可以处理的数据集的大小。

2) 由于采用了预排序技术，而排序算法的复杂度本身并不是与记录个数成线性关系，因此，使得 SLIQ 算法不可能达到随记录数目增长的线性可伸缩性。

(4) SPRINT 算法

为了减少驻留于内存的数据量,SPRINT 算法进一步改进了决策树算法的数据结构，去掉了在 SLIQ 中需要驻留于内存的类别列表，将它的类别列合并到每个属性列表中。这样，在遍历每个属性列表寻找当前节点的最优分裂标准时，不必参照其他信息，将对节点的分裂表现在对属性列表的分裂，即将每个属性列表分成两个，分别存放属于各个节点的记录。

SPRINT 算法的优点是在寻找每个节点的最优分裂标准时变得更简单。其缺点是对非分裂属性的属性列表进行分裂变得很困难。解决的办法是对分裂属性进行分裂时用哈希表记录下每个记录属于哪个孩子节点,若内存能够容纳下整个哈希表，其他属性列表的分裂只需参照该哈希表即可。由于哈希表的大小与训练集的大小成正比，当训练集很大时，哈希表可能无法在内存容纳，此时分裂只能分批执行，这使得 SPRINT 算法的可伸缩性仍然不是很好。

此外，基于决策树的主要改进算法还包括 CART（classification and regression tree)、PUBLIC(pruning and building integreated in classification)等。

8.3.2 贝叶斯分类

贝叶斯分类是统计学分类方法，它是一类利用概率统计知识进行分类的算法。它在先验概率与条件概率已知的情况下,预测类成员关系可能性的模式分类算法。如计算一个给定样本属于一个特性类的概率,并选定其中概率最大的一个类别作为该样本的最终判别。假设每个训练样本用一个 n 维特征向量 $X=\{x_1, x_2, \cdots, x_n\}$

表示，分别描述 n 个属性 A_1，A_2,…,A_n对样本的测量。将训练样本集分为 m 类，记为 C_1，C_2,…,C_m。贝叶斯原理通常用下面的公式来表示：

$$P(C_i|X) = \frac{P(x|Ci)P(Ci)}{\sum_{j=1}^{m} P(x|Cj)P(Cj)}$$

其中，X 表示观测数据样本，C_j为某种假设，$P(C_i)$是 C_i的先验概率，（i, j=1, 2, .., m）。P(X| Ci)是条件概率，先验概率对条件概率加权平均后，得到条件 X 下，C_i的后验概率 $P(C_i|X)$。上述是朴素贝叶斯的工作过程，也是贝叶斯分类算法的判别准则。

在许多场合，朴素贝叶斯（Naïve Bayes, NB）分类可以与决策树和神经网络分类算法相媲美，该算法能运用到大型数据库中，且方法简单、分类准确率高、速度快。由于贝叶斯定理假设一个属性值对给定类的影响独立于其它的属性值，而此假设在实际情况中经常是不成立的，因此其分类准确率可能会下降。为此，就出现了许多降低独立性假设的贝叶斯分类算法，如 TAN(tree augmented Bayes network)算法、贝叶斯网络分类器（Bayesian network classifier，BNC）。

（1）朴素贝叶斯算法

朴素贝叶斯分类器以简单的结构和良好的性能受到人们的关注，它是最优秀的分类器之一。朴素贝叶斯分类器建立在一个类条件独立性假设(朴素假设)基础之上：给定类节点(变量)后，各属性节点(变量)之间相互独立。朴素贝叶斯分类器可以看作是贝叶斯网络的一种最简化的模型。

根据朴素贝叶斯的类条件独立假设，则有：

$$P(X|C_i)=\prod_{k=1}^{m} P(X_K|C_i)$$

条件概率 $P(X_1|C_i)$，$P(X_2|C_i)$,…,$P(X_n|C_i)$可以从训练数据集求得。根据此方法，对一个未知类别的样本 X，可以先分别计算出 X 属于每一个类别 C_i的概率 $P(X|C_i)P(C_i)$,然后选择其中概率最大的类别作为其类别。

朴素贝叶斯算法成立的前提是各属性之间相互独立。当数据集满足这种独立性假设时，分类的准确度较高，否则可能较低。另外，该算法没有分类规则输出。

（2）TAN 算法

TAN 算法通过发现属性对之间的依赖关系来降低 NB 中任意属性之间独立的假设。它是在 NB 网络结构的基础上增加属性对之间的关联（边）来实现的。

实现方法是：用节点表示属性，用有向边表示属性之间的依赖关系，把类别属性作为根节点，其余所有属性都作为它的子节点。通常，用虚线代表 NB 所需

的边，用实线代表新增的边。属性 A_i 和 A_j 之间的边意味着属性 A_i 对类别变量 C 的影响还取决于属性 A_j 的值。这些增加的边满足下列条件：类别变量没有双亲节点，每个属性有一个列别变量双亲节点和最多另外一个属性作为其双亲节点。找到这组关联边之后，就可以计算一组随机变量的联合概率分布如下：

$$P(A_1, A_2, \cdots, A_n) = P(C)\prod_{i=1}^{n} P(Ai|\prod Ai)$$

其中$\square\prod Ai$代表的是 Ai 的双亲节点。由于在 TAN 算法中考虑了 n 个属性之间独立性的假设有了一定程度的降低，但是属性之间可能存在更多其它的关联性仍没有考虑，因此其使用范围仍然受到限制。

（3）贝叶斯网络分类器

贝叶斯网络分类器放弃了朴素贝叶斯分类器的条件独立性假设，所以最能与领域数据相吻合。在贝叶斯网络的结构中类节点地位同其他属性节点一样，也可以有父节点。本文采用基于搜索打分的方法构造贝叶斯分类器，搜索打分算法采用 K2 搜索算法和 BIC 评分函数。

贝叶斯网络分类方法如下：

1）输入：训练集 D；变量顺序；变量父节点个数上界 u；

2）K2 算法构造 BNC:

a、所有节点组成无向图

b、确定变量 j_x 的父节点个数，等于 u 则停止为它寻找父节点；

c、如果父节点的个数大于 u，则从中按顺序选择 j_x 之前的节点，但不是 j_x 父节点的变量 i_x 做为 j_x 的父节点；

d、使用 BIC 测度对新结构打分；

e、同前次打分比较，如果评分高，则添加 i_x 为 j_x 的父节点；如果 BIC 评分低，则停止为 j_x 寻找父节点；

3）使用训练数据集进行参数学习（最大似然估计法）；

4）对测试集分类，得出分类准确度。

下面主要从分类准确度和分类耗时这两个方面分析比较这三种分类器。

1）朴素贝叶斯分类器。从分类准确度上看，NBC 虽然结构简单但是它的分类准确度并不低。从分类耗时看，NBC 普遍比其它两种分类器花费的时间少，这与它不需要结构学习，计算复杂度低是密切相关的。NBC 在现实中有着广泛的适应

性，这主要还因为在大部分领域中属性之间的依赖关系要明显低于属性和类别之间的依赖关系，所以 NBC 的条件独立性假设是具有一定的现实意义的。

2）基于 BIC 测度的 TAN 分类器是所有 NBC 改进分类器中效果最好的一个。TAN 分类器的分类准确度普遍高于 NBC，TAN 分类器放松了条件独立性假设这是同现实世界相符合的，当属性之间关联性越大时，TAN 分类器的效果就越好。TAN 分类器中需要设置根节点，根节点就是选择除去类节点以外的属性节点作为其它属性节点的根节点，根节点的设置对分类准确度并没有很大的影响。从分类时间上看，TAN 分类器在这三种分类器中是花费时间最长的。

3）理论上 BNC 分类器应该有最好的分类效果，但是实际中，BNC 的分类效果并不理想，这主要与两个因素有关：一是数据集的规模。BNC 对大样本的数据集有较好的分类效果，在小规模数据集情况下就不如 NBC 和 TAN；二是在使用 K2 算法进行结构学习的过程中有一个重要的参数，用来确定结点变量的次序，它对先验知识的依赖性很大。在不了解相关的领域或没有专家的指导的情况下，确定变量的次序就变得相当困难。

从分类耗时上看，BNC 分类器的分类耗时比 NBC 要长，同 TAN 比较有一定的不确定性，它普遍要比 TAN 分类时间短。这三种分类器并不是对每种数据集都有好的分类效果，因此在对数据集选择分类器的时候还需要具体情况具体对待，主要考查属性之间的关联性、数据的规模和时间限制等方面。数据集属性相关性小的时候选择 NBC 有较好的分类效果，数据集属性相关性大时候选择 TAN 分类器。在数据集规模较大且具有一定先验知识时选择贝叶斯网络分类器。

8.3.3 k-近邻

k-近邻(kNN，k-Nearest Neighbors)算法是一种基于实例的分类方法，是一种非参数的分类技术。

基本原理：kNN 分类算法搜索样本空间，计算未知类别向量与样本集中每个向量的相似度，在样本集中找出 K 个最相似的文本向量，分类结果为相似样本中最多的一类。

但在大样本集和高维样本分类中（如文本分类），kNN 方法的缺陷凸显。表现在以下几个方面：1）KNN 分类算法是懒散的分类算法，对于分类所需的计算均推迟至分类进行，故在其分类器中存储有大量的样本向量。在未知类别样本需

要分类时，在计算所以存储样本和未知类别样本的距离时，高维样本或大样本集所需要的时间和空间的复杂度均较高。2）KNN 分类算法是建立在 VSM 模型上的，其样本距离测度使用欧氏距离。若各维权值相同，即认定各维对于分类的贡献度相同，显然这不符合实际情况。

基于上述缺点，人们也采用了一些改进算法：当样本数量较大时，为减小计算，可对样本集进行编辑处理，即从原始样本集中选择最优的参考子集惊醒 KNN 计算，以减少样本的存储量和提高计算效率。

8.3.4 基于数据库技术的分类算法

虽然数据挖掘的创始人主要是数据库领域的研究人员，然而提出的大多数算法则没有利用数据库的相关技术。在分类算法中，致力于解决此问题的算法有

MIND(mining in database)和 GAC-RDB(grouping and counting-relational database)。

（1）MIND 算法

MIND 算法是采用数据库中用户自定义的函数(user-defined function, UDF)实现发现分类规则的算法。MIND 采用典型的决策树构造方法构建分类器。具体步骤与 SLIQ 类似。其主要区别在于它采用数据库提供的 UDF 方法和 SQL 语句实现树的构造。简而言之，就是在树的每一层，为每一个属性建立一个维表，存放各属性的每个取值属于各个类别的个数以及所属的结点编号。根据这些信息可以为当前结点计算每种分裂标准的值，选出最优的分裂标准，然后据此对结点进行分裂，修改维表中结点编号列的值。在上述过程中，对维表的创建和修改需要进行多次，若用 SQL 实现，耗时很多，因此用 UDF 实现。而分类标准的寻找过程则通过创建若干表和视图，利用连接查询实现。

该算法的优点是通过采用 UDF 实现决策树的构造过程使得分类算法易于与数据库系统集成。其缺点是算法用 UDF 完成主要的计算任务，而 UDF 一般是由用户利用高级语言实现的，无法使用数据库系统提供的查询处理机制，无法利用查询优化方法，且 UDF 的编写和维护相当复杂。此外，MIND 中用 SQL 语句实现的那部分功能本身就是比较简单的操作，而采用 SQL 实现的方法却显得相当复杂。

（2）GAC-RDB 算法

GAC-RDB 算法是一种利用 SQL 语句实现的分类算法。该算法采用一种基于分组计数的方法统计训练数据集中各个属性取值组合的类别分布信息，通过最小置信度和最小支持度两个阈值找出有意义的分类规则。在该算法中，首先利用 SQL 语句计算每个属性进行类别判定的信息量，从而选择一个最优的分裂属性，并且按照信息量的大小对属性进行排序，然后重复地进行属性的选择、候选分类表的生成、剪裁以及分类误差的计算，直到满足结束条件为止。比如，直到小于误差阈值和误差没有改变为止。

该算法的优点是具有与现有的其他分类器相同的分类准确度，执行速度有较大提高，而且具有良好的伸缩性，应用程序易于与数据库系统集成。其缺点是参数的取值需用户完成等。

8.3.5 基于关联规则的分类算法

关联规则挖掘是数据挖掘研究的一个重要的、高度活跃的领域。近年来，数据挖掘技术已将关联规则挖掘用于分类问题，取得了很好的效果。

ARCS(Association Rule Clustering System)基于聚类挖掘关联规则，然后使用规则进行分类。将关联规则画在 2-D 栅格上，算法扫描栅格，搜索规则的矩形聚类。实践发现，当数据中存在孤立点时，ARCS 比 C4.5 稍微精确一点。ARCS 的准确性与离散化程度有关。从可伸缩性来说，不论数据库多大，ARCS 需要的存储容量为常数。

CBA(classification based on association)是基于关联规则发现方法的分类算法。该算法分两个步骤构造分类器。第一步：发现所有形如 $x_{i1} \wedge x \Rightarrow Ci$ 的关联规则，即右部为类别属性值的类别关联规则(classification association rules，CAR)。第二步：从已发现的 CAR 中选择高优先度的规则来覆盖训练集，也就是说，如果有多条关联规则的左部相同，而右部为不同的类，则选择具有最高置信度的规则作为可能规则。文献[4]对该过程进行了较深入的研究，使得算法在此步骤不需要对训练数据集进行过多的扫描。

CBA 算法的优点是其分类准确度较高，在许多数据集上比 C4.5 更精确。此外，上述两步都具有线性可伸缩性。

CBA(Classification Based on Association)是关联分类。此算法把分类规则挖掘和关联规则挖掘整合到一起。与 CART 和 C4.5 只产生部分规则不同的

是,CBA 产生所有的类关联规则 CARs(Class Association Rules)，然后选择最好的规则去覆盖训练集。另外，在此算法的框架中，数据库可以驻留在磁盘中

CAEP 使用项集支持度挖掘 HV 露模式(Emerging Pattern)，而 EP 用于构造分类。CAEP 找出满足给定支持度和增长率阈值的 EP。已经发现，在许多数据集上，CAEP 比 C4.5 和基于关联的分类更精确。一种替代的、基于跳跃的 HV 露模式 JEP(Jnmping Emerging Pattern)是一种特殊类型的 EP，项集的支持度由在一个数据集中的 0 陡峭地增长到另一个数据集中的非 0。在一此大的多维数据库中，JEP 性能优于 CAEP，但在一些小型数据库中，CAEP 比 JEP 优，这二种分类法被认为是互补的。

ADT(Association Decision Trec)分二步实现以精确度驱动为基础的过度适合规则的剪枝。第一步，运用置信度规则建立分类器。主要是采用某种置信度的单调性建立基于置信度的剪枝策略。第二步，为实现精确性，用关联规则建立一种平衡于 DT(Dccision Tree)归纳的精确度驱动剪枝。这样的结果就是 ADT(Association Based Decision Trec)。它联合了大量的关联规则和 DT 归纳精确性驱动剪枝技术。

基于多维关联规则的分类算法 CMAR(Classification Based on Multiple Class-Association Rules)是利用 FP-Growth 算法挖掘关联规则，建立类关联分布树 FP－树。采用 CR－树(Classification Rulc Trcc)结构有效地存储关联规则。基于置信度、相关性和数据库覆盖来剪枝。分类的具体执行采用加权厂来分析。与 CBA 和 C 4.5 相比，CMAR 性能优异且伸缩性较好。但 CMAR 优先生成的是长规则，对数据库的覆盖效果较差;利用加权 x2 统计量进行分类，会造成 x2 统计量的失真，致使分类值的准确程度降低。CPAR(Classification Based on Predictive Association Rules)整合了关联规则分类和传统的基于规则分类的优点。为避免过度适合，在规则生成时采用贪心算法，这比产生所有候选项集的效率高;采用一种动态方法避免在规则生成时的重复计算;采用顶期精确性评价规则，并在预测时应用最优的规则，避免产生冗余的规则。另外，MSR(Minimnm Set Rule)针对基于关联规则分类算法中产生的关联规则集可能太大的问题，在分类中运用最小关联规则集。在此算法中，CARS 并不是通过置信度首先排序，因为高置信度规则对噪声是很敏感的。采用早期剪枝力方法可减少关联规则的数量，

并保证在最小集中没有不相关的规则。实验证实，MSR 比 C45 和 CBA 的错误率要低得多。

8.3.6 支持向量机分类

支持向量机（Support Vector Machine）是 Cortes 和 Vapnik 与 1995 年首先提出的，它在解决小样本、非线性及高维模式识别中有许多特有的优势，并能推广应用到函数拟合等其他机器学习问题中。

支持向量机（SVM）方法是建立在统计学习理论的 VC 维和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以期获得最好的推广能力。SVM 是从线性可分情况下的最优分类面发展而来的，使分类间隔最大实际上就是对推广能力的控制，这是 SVM 的核心思想之一。

由于统计学习理论和支持向量机建立了一套较好的在小样本下机器学习的理论框架和通用方法，既有严格的理论基础，又能较好地解决小样本、高维和局部极小点等实际问题，因此成为继神经网络之后的又一个研究方向。

但是，处理大规模数据集时，SVM 速度慢，往往需要较长的训练时间。而且，SVM 方法需要计算和存储核函数矩阵，当样本数目较大时，需要很大的内存。其次，SVM 在二次型寻优过程中要进行大量的矩阵运算，多数情况下，寻优算法是占用算法时间的主要部分。

8.3.7 人工神经网络算法

神经网络是分类技术中重要方法之一。人工神经网络（Artificial Neural Networks，ANN）是一种应用类似于大脑神经突触联接的结构进行信息处理的数学模型。在这种模型中，大量的节点（或称”神经元”，或”单元”）之间相互联接构成网络，即”神经网络”，以达到处理信息的目的。神经网络通常需要进行训练，训练的过程就是网络进行学习的过程。训练改变了网络节点的连接权的值使其具有分类的功能，经过训练的网络就可用于对象的识别。神经网络的优势在于：（1）可以任意精度逼近任意函数；（2）神经网络方法本身属于非线性模型，能够适应各种复杂的数据关系；（3）神经网路具备很强的学习能力，使它能比很多分类算法更好地适应数据空间的变化；（4）神经网络借鉴人脑的物理结构和机理，能够模拟人脑的某些功能，具备“智能”的特点。

用于分类常见的神经网络模型包括：BP（Back Propagation）神经网络、RBF网络、Hopfield网络、自组织特征映射神经网络、学习矢量化神经网络。目前神经网络分类算法研究较多集中在以BP为代表的神经网络上。当前的神经网络仍普遍存在收敛速度慢、计算量大、训练时间长和不可解释等缺点。

（1）BP神经网络分类算法

BP神经网络是一种多层前馈神经网络，该网络的主要特点是信号前向传递，误差反向传播。它是目前应用最广泛的一种前向神经网络模型。在前向传递中，输入信号从输入层经隐含层逐层处理，直至输出层。每一层的神经元状态只影响下一层神经元状态。如果输出层得不到期望输出，则转入反向传播，根据预测误差调整网络权值和阈值，从而使BP神经网络预测输出不断逼近期望输出。

BP神经网络存在一些缺陷,它值适用于平稳环境,学习算法计算的费用较高，不具备自学能力，不能进行快速学习、记忆以及学习能力之间存在冲突等问题，虽然有多种改进算法，但仍不能从根本上解决这些问题。另外，此类神经网络借鉴了人脑的物理结构，存储在神经网络中的知识往往以权值的形式表现出来，这种形式本身很难理解。

（2）RBF神经网络

径向基函数（RBF,Radical Basis Function）是多维空间插值的传统技术，有Powell于1985年提出。1988年，Broomhead和Lowe根据生物神经元具有局部响应这一特点，将RBF引入神经网络设计中，产生了RBF神经网络。1989年，Jackson论证了RBF神经网路对非线性连续函数的一致逼近性能。

RBF神经网络属于前向神经网络类型,网络的结构与多层前向神经网络类似,是一种三层的前向网络,第一层为输入层,有信号源结点组成;一二层为隐含层,隐含层节点数视所描述问题的需要而定,隐藏层中神经元的变换函数及径向基函数是对中心点径向对称且衰减的非负非线性函数，该函数是局部响应函数；第三层为输出层，它对输入模式作出响应。

RBF网络的基本思想:用RBF作为隐单元的“基”构成隐含层空间，隐含层对输入矢量进行变换，将低维的模式输入数据变换到高维空间内，使得在低维空间内的线性不可分的问题在高维空间内线性可分。

（3）SOFM神经网络

受生物系统视网膜皮层生物特性和大脑皮层区域“有序特征映射”的影响，Kohonen 提出了自组织特征映射神经网络（SOFM），这种网络在网络输出层具备按照几何中心或者特征进行聚合的独特特质。它由输入层和竞争层构成，竞争层有一维或者二维阵列的神经元组成，输入层和竞争层之间实现全连接。通过在竞争学习过程中动态改变活性泡大小，该结构具备拓扑结构保持、概率分布保持、克石化等诸多优点。SOFM 神经网络竞争层神经元个数要求事先指定，这种限制极大地影响了其在实际中的使用。针对此不足人们又提出了动态自组织特征映射神经网络，最具有代表性的是 D.Alahakoon 等提出的 GSOFM(growing self-organizing maps)模型。

（4）学习矢量化（LVQ）神经网络

该网络是对 Kohonen 神经网络的监督学习的扩展形式，允许对输入分类进行指定。学习矢量化神经网络有输入层、竞争层、线性层构成。线性层神经元代表不同类别，竞争层的每一个神经元代表每个类别中的一个子类；线性层和竞争层之间用矩阵实现子类和类之间的映射关系。竞争层和输入层之间是类似于 SOFM 神经网络的结构。LVQ 神经网络已 LVQ 为基本模型，一次为基础提出改进模型 LVQ2 和 LVQ3。这三者之间的不同点在于，早 LVQ 中只有获胜神经元才会得到训练，而在 LVQ2 和 LVQ3 中，当适当条件符合时，学习矢量化可以通过训练获胜神经元和次获胜神经元来对 SOFM 网络的训练规则进行扩展。

人工神经网络作为另一种处理非线性、不确定性的有力工具，目前还存在许多局限性。首先，网络本身的黑箱式内部知识表达，使其不能利用初始经验进行学习，易于陷入局部极小值。其次，就本质而言，人工神经网络就是用静态网络处理连续时间动态系统的控制问题。这就不可避免地带来了差分模型定阶及网络规模随阶次迅速增加的复杂性问题。再次，人工神经网络的泛化能力在相当程度上决定了控制系统的鲁棒性。全局逼近的泛化能力受大量局部极值与缓慢学习速度制约，局部逼近则受存储容量与实时性的而严重限制。

8.3.8 基于软计算的分类方法

在数据挖掘领域，软计算的用途越来越广泛：模糊逻辑用于处理不完整、不精确的数据以及近似答案等；神经网络用于高分线性决策、泛化学习、自适应、自组织和模式识别；遗传算法用于动态环境下的高效搜索、复杂目标对象的自适

应和优化；粗糙集根据“核”属性获得对象的近似描述，能有效处理不精确、不一致、不完整等各种不完备信息。当数据集表现出越来越多的无标签性、不确定性、不完整性、非均匀性和动态性特点时，传统数据挖掘算法对此往往无能为力，软计算却为此提供一种灵活处理数据的能力，软计算内的融合与传统数据挖掘方法的结合逐渐成为数据挖掘领域的研究趋势。

（1）粗糙集（rough set）

粗糙集理论是一种刻划不完整和不确定性数据的数学工具，不需要先验知识，能有效地处理各种不完备信息，从中发现隐含的知识，并和各种分类技术相结合建立起来能够对不完备数据进行分类的算法。粗糙集理论将分类能力和知识联系在一起，使用等价关系来形式化地表示分类，知识因而表示为等价关系集 R 对离散空间 U 的划分。粗糙集理论还包括求取数据中最小不变集合最小规则集的额理论，即简约算法（即分类中属性简约和规则生成），其基本原理是通过求属性的重要性并排序，在泛化关系中找出与原始数据具有同一决策或分辨能力的相关属性的最小集合，以此实现信息简约，这也是粗糙集理论在分类中的应用。

简约主要借助于信息表达这样一种有效的知识表达形式；在保持信息表中决策属性和条件属性依赖关系不变时进行的信息表简约，具体包括属性简约和值简约。

属性简约在一定程度上对信息表中的非必要的冗余信息进行简约，但对每一个实例而言仍可能存在不必要的属性，因此在不引起冲突的条件下可将每一个实例的不必要属性删除，即为值简约。值简约的最终结果就是分类所需要的规则，常见的值简约算法包括归纳值简约、启发式值简约、基于决策矩阵的值简约算法、增量式规则获取算法和其他一些改进算法。

粗糙集本身限制条件较强，在实际应用中，可以模态逻辑和概率统计信息对基本粗糙集模型进行扩展，从而形成了代数粗糙集模型和概率统计粗糙集模型。

（2）遗传算法

遗传算法在解决多峰值、非线性、全局优化等高复杂度问题时具备独特优势，它是以基于进化论原理发展起来的高效随机搜索与优化方法。它以适应函数为依据，通过对群体、个体施加遗传操作来实现群体内个体结构的优化重组。在全局范围内逼近最优解。遗传算法综合了定向搜索与随机搜索的优点。避免了大多数经典优化方法基于目标函数的梯度或高阶导数而易陷入局部最优的缺陷，可以取

得较好的区域搜索与空间扩展的平衡。在运算时随机的多样性群体和交叉运算利于扩展搜索空间；随着高适应值得获得，交叉运算利于在这些解周围搜索。遗传算法由于通过保持一个潜在解的群体进行多方向的搜索而有能力跳出局部最优解。

遗传算法的应用主要集中在分类算法等方面。而基本思路如下：数据分类问题看成是在搜索问题，数据库看作是搜索空间，分类算法看作搜索策略。因此，应用遗传算法在数据库中进行搜索，对随机产生的一组分类规则进行进化，直到数据库能被该组分类规则覆盖，从而挖掘出隐含在数据库中的分类规则。应用遗传算法进行数据分类，首先要对实际问题进行编码；然后定义遗传算法的适应度函数，由于算法用于规则归纳，因此，适应度函数有规则覆盖的正例和反例来定义。1978年Holland实现了第一个基于遗传算法的机器学习系统CS-1(cognitive system level one)，其后又提出了桶队（Bucket Brigade）算法。1981年Smith实现了与CS-1有重大区别的分类器LS-1, 以此为基础，人们又提出了基于遗传算法的分类系统，如GCLS(genetic classifier learning system)等算法。

（3）模糊逻辑

模糊数学是研究模糊现象数学。模糊数学最基本概念是隶属函数，即以一个值域在[0,1]之间的隶属函数来描述论域中对象属于某一个类别的程度，并以此为基础确定模糊集的运算法则、性质、分解和扩展原理、算子、模糊的、模糊集的近似程度等度量概念和算法。

分类操作离不开向量相似程度的计算，而模糊分类操作也需要向量模糊相似系数的计算。在模糊分类方法中，首先要建立模糊相似矩阵，表示对象的模糊相似程度其元素称为模糊相似系数，其确定方法主要有：数量积法、夹角余弦法、相关系数法、最大最小法、算术平均最小法、几何平均最小法、绝对值指数法、指数相似系数法、绝对值倒数法、绝对值减数法、参数法、贴近度法。

模糊分类方法可以很好地处理客观事物类别属性的不明确性，主要包括传达闭包法、最大树法、编网法、基于摄动的模糊方法等。但人们更多地将模糊方法与其他分类算法结合起来，既有与传统分类算法，如模糊决策树、模糊关联规则挖掘等的结合，也有与软计算在内其他算法，如模糊神经网络等的结合。

8.3.9 其他分类算法

（1） LB 算法

LB(Large Bayes)算法是一种基于概率统计和关联规则的分类算法。在算法的训练阶段，利用挖掘关联规则的 Apriori 算法找出训练集中所有的频繁且有意义的项目集，存放在集合 F 中。对于一个未知类别的样本 A，可以从 F 中找出包含在 A 中的最长的项目集来计算 A 属于各个类别的概率，并且选择其中概率最大的类别为其分类。LB 算法的分类准确度比现有的其他分类算法的准确度好。但该算法仍有与贝叶斯算法和 CBA 算法相同的缺点。

（2） CAEP 算法

CAEP(classification by aggregating emerging patterns)算法使用项目集支持度挖掘显露模式（emerging pattern，EP），再用 EP 构造分类器。一个 EP 是一个项目集，其支持度由一个类别到另一个类别显著增加，两个支持度的比称作 EP 的增长率。例如，假定有顾客数据集，包含两个类 C1 和 C2，分别代表 buy-computer=yes 和 buy-computer=no。若项目集 age 30，student=no 是一个 EP，其支持度由在 C1 中的 0.2%增加到 C2 中的 57.6%，则增长率为 57.6%/0.2%=288。如果一个新样本 X 包含在该 EP 中，则可以说 X 属于 C2 的几率为 99.65%。

许多数据集 CAPE 比 C4.5 和基于关联的分类更精确。

8.4 分布式分类

Mahout 中的分类算法可以广泛地应用于海量数据分类工程中。一般来说，Mahout 分类算法对资源的要求不会快于训练数据和测试数据的增长速度，而且能够转换为分布式处理，从而可以通过扩展分布式处理节点的数量解决大规模数据集分类问题。如图 14 所示，当训练例子的数量相对较小时，与 mahout 分类算法相比，传统数据挖掘方法能有相同或更好的性能。但当样本数量增加后，传统不可扩展的分类算法架构所需的处理时间快速增加，此时 Mahout 的可伸缩和并行算法的优势就变得明显。

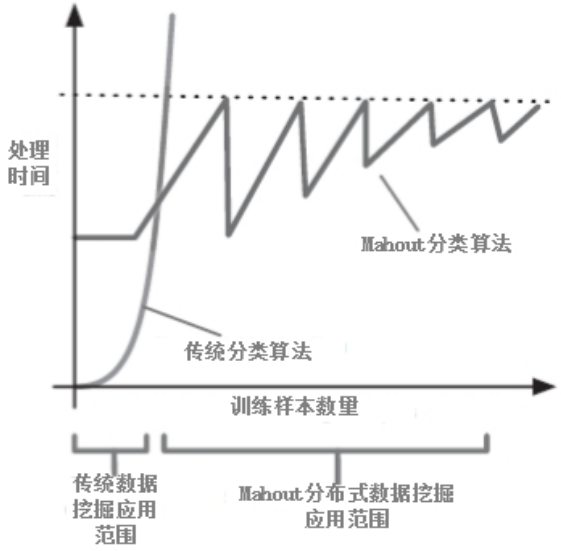


图 14 使用 Mahout 分类和传统分类算法的处理时间

Mahout 的主要优势是其强劲的分布式处理能力，可应对不断增长的数据集。Mahout 拥有多种分类算法，其中大多数算法都是为了在 Hadoop 上运行而编写的（除了随机梯度下降法），并具有可伸缩性。表 9 列举了 Mahout 中的各种分类算法和其所适用的数据规模，并对于特定分类问题推荐了最合适的算法。以朴素贝叶斯算法为例说明 Mahout 中的分类算法处理流程。朴素贝叶斯分类算法以速度快和准确性高而著称，但其关于数据的简单（通常也是不正确的）假设是完全独立的，该算法包括两个流程：第一个步骤称作训练（training），它将通过特定的分类算法学习已标记类别的样本来创建一个分类器，并使用部分已标记类别的样本测试该分类器的准确性。第二个步骤称作分类，它将使用在训练阶段中创建的分类器对未标记类别的样本进行归类，实现对于样本类别的预测。因此，要运行 Mahout 的分类器，首先需要训练得到分类器，然后再使用该分类器对新内容进行分类。

表格 9 Mahout 分类学习算法特点

数据集类型	Mahout 算法	处理模式	特点
小型到中型数据集（千万等级以下的训练样本数目）	随机梯度下降算法 (Stochastic gradient descent，SGD)	序列化，增量式	适用于所有数据类型的分类和预测，在百万训练样本等级上运行效率高。
中型到大型数据集（千万等级到数亿等级的训练样本数目）	支持向量机算法 (Support Vector Machine， SVM)	序列化	仍在实验阶段，可在大型数据集中使用
	朴素贝叶斯分类算法 (Naive Bayes)	分布式	非常适合对文本数据进行分类，中到高等级的训练开销，可有效处理对于 SGD 算法来说太大的数据集

	互补朴素贝叶斯分类算法(Complementary Naive Bayes)	分布式	有效处理对于 SGD 算法来说太大的数据集，并解决了朴素贝叶斯算法的使用限值（如数据的独立性不符合要求）。
小型到中型数据集（千万等级以下的训练样本数目）	随机森林算法(Random forests)	分布式	适用于所有数据类型的分类和预测，训练开销高，适用于复杂的分类场景，能处理数据中的非线性和条件关系

Mahout 中朴素贝叶斯和互补朴素贝叶斯算法均可并行运行，在多个有效机器中工作，因此他们可应用于比 SGD 算法更大的数据集中。如果要解决千万级别的训练样本，并且预测变量是单一的、文字性的值，朴素贝叶斯或互补朴素贝叶斯算法是最佳选择。Apache Mahout 官方网站给出了两个非常适合朴素贝叶斯算法解决的实例：Wikipedia Example 和 Twenty Newsgroups Classification Example，可以通过运行这两个案例来了解 Mahout 分类算法的运行过程。以 Wikipedia Example 为例，在运行训练程序和分类器之前，首先需要下载用于训练和测试的文档样本。在获取 Wikipedia 副本的基础上有 HDFS 将数据划分为存储块，然后对每个存储块内的数据按照国家运行分类算法，训练一个分类器来预测一篇文章所属的国家类别，代表此国家对于该文章没有阅读权限。

8.5 总结

分类是数据挖掘的重要方法之一。到目前为止，基于各种思想和理论基础的分类算法被提出，算法的实际应用也已趋于成熟。但实践证明，没有一种分类算法对所有的数据类型都优于其他分类算法。每种相对较优的算法都有它具体的应用环境。以上简单介绍了各种主要的分类方法，应该说其都有各自不同的特点及优缺点。对于数据库负载的自动识别，应该选择哪种方法呢？用来比较和评估分类方法的标准主要有：（1）预测的准确率。模型正确地预测新样本的类标号的能力；（2）计算速度。包括构造模型以及使用模型进行分类的时间；（3）强壮性。模型对噪声数据或空缺值数据正确预测的能力；（4）可伸缩性。对于数据量很大的数据集，有效构造模型的能力；（5）模型描述的简洁性和可解释性。模型描述愈简洁、愈容易理解，则愈受欢迎。

未来数据分类算法的研究则更多地集中在智能分类领域，如基于软计算的分类算法以及免疫算法、分形编码、蚁群优化等智能算法的分类研究上。

9 面向未知分布：聚类分析

9.1 数据分布与聚类

聚类(Clustering)分析是数据挖掘技术的重要组成部分,它能从潜在的数据中发现新的、有意义的数据分布模式,已经广泛应用于模式识别、数据分析、图像识别及其他许多方面。聚类[60]是在事先不规定分组规则的情况下,将数据按照其自身特征划分成不同的群组。其重要特征是“物以类聚”,即要求在不同群组的数据之间差距越大、越明显越好,而每个群组内部的数据之间要尽量相似,差距越小越好。

聚类是一个具有挑战性的研究领域,目前对聚类算法的研究非常多。基本上所有的聚类算法都具有其各自的特点,只适用于某些特定领域,目前还没有能适用于各种领域的聚类算法。如较常用的 K-MEANS 算法主要以方法简单、执行效率高见长,但只能识别大小近似的球形类; DBSCAN 算法能很好地过滤噪声数据,但其时间复杂度却为 $O(n^2)$,效率不高。聚类算法大体可分为五类:划分方法、层次方法、基于密度的方法、基于网格的方法以及基于模型的方法。

9.2 聚类算法的种类及特性

聚类是一种常见的数据分析工具,其目的是把大量数据点的集合分成若干类,使得每个类中的数据之间最大程度地相似,而不同类中的数据最大程度地不同。在多媒体信息检索及数据挖掘的过程中,聚类处理对于建立高效的数据库索引、实现快速准确的信息检索具有重要的理论和现实意义。聚类算法按所采用的基本思想为依据可以分为五类,即层次聚类算法、分割聚类算法、基于约束的聚类算法、机器学习中的聚类算法以及用于高维数据的聚类算法,如图 15 所示。

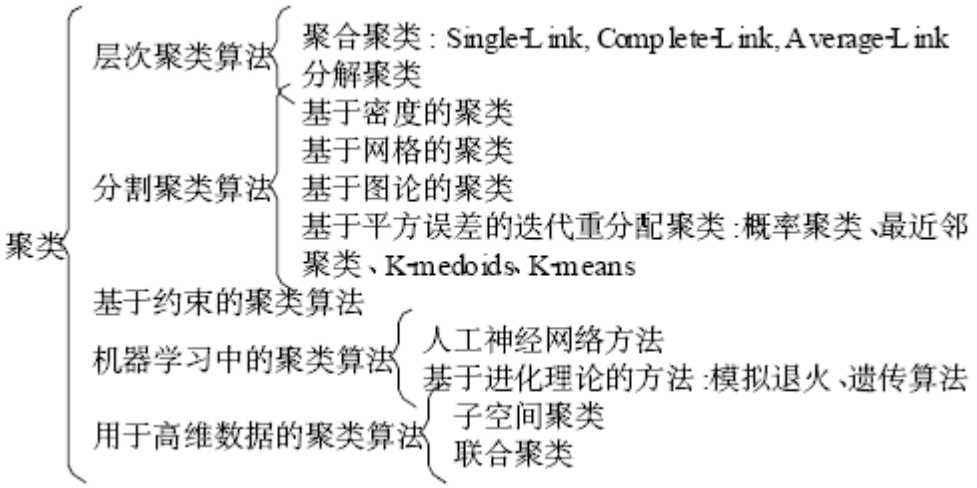


图 15 聚类算法分类示意图

9.2.1 层次聚类算法

层次聚类算法通过将数据组织成若干组并形成一个相应的树状图来进行聚类,它又可以分为两类,即自底向上的聚合层次聚类和自顶向下的分解层次聚类。聚合聚类的策略是先将每个对象各自作为一个原子聚类,然后对这些原子聚类逐层进行聚合,直至满足一定的终止条件;后者则与前者相反,它先将所有的对象都看成一个聚类,然后将其不断分解直至满足终止条件。

对于聚合聚类算法来讲,根据度量两个子类的相似度时所依据的距离不同,又可将其分为基于 Single-Link, Complete-Link 和 Average-Link 的聚合聚类。Single-Link 在这三者中应用最为广泛,它根据两个聚类中相隔最近的两个点之间的距离来评价这两个类之间的相似程度,而后两者则分别依据两类中数据点之间的最远距离和平均距离来进行相似度评价。

CURE, ROCK 和 CHAMELEON 算法是聚合聚类中最具代表性的三个方法。

Guha 等人在 1998 年提出了 CURE 算法[61]。该方法不用单个中心或对象来代表一个聚类,而是选择数据空间中固定数目的、具有代表性的一些点共同来代表相应的类,这样就可以识别具有复杂形状和不同大小的聚类,从而能很好地过滤孤立点。ROCK 算法[62]是对 CURE 的改进,除了具有 CURE 算法的一些优良特性之外,它还适用于类别属性的数据。CHAMELEON 算法[63]是 Karyp is 等人于 1999 年提出来的,它在聚合聚类的过程中利用了动态建模的技术。

9.2.2 分割聚类算法

分割聚类算法是另外一种重要的聚类方法。它先将数据点集分为 k 个划分,然后从这 k 个初始划分开始,通过重复的控制策略使某个准则最优化以达到最终的结果。这类方法又可分为基于密度的聚类、基于网格的聚类、基于图论的聚类和基于平方误差的迭代重分配聚类。

（1）基于密度的聚类

基于密度的聚类算法从数据对象的分布密度出发,将密度足够大的相邻区域连接起来,从而可以发现具有任意形状的聚类,并能有效处理异常数据。它主要用于对空间数据的聚类。

DBSCAN[64]是一个典型的基于密度的聚类方法,它将聚类定义为一组密度连接的点集,然后通过不断生长足够高密度的区域来进行聚类。DENCLUE[65]则根据数据点在属性空间中的密度来进行聚类。从本质上讲,DENCLUE 是基于密度的聚类算法与基于网格的预处理的结合,它受目标数据的维度影响较小。此外,Ankerst 等人提出的 OPTICS, Xu 等人提出的 DBCLASD 和马帅等人提出的 CURD 算法也采用了基于密度的聚类思想,它们均针对数据在空间中呈现的不同密度分布对 DB2SCAN 作了相应的改进。

（2）基于网格的聚类

基于网格的聚类从对数据空间划分的角度出发,利用属性空间的多维网格数据结构,将空间划分为有限数目的单元以构成一个可以进行聚类分析的网格结构。该方法的主要特点是处理时间与数据对象的数目无关,但与每维空间所划分的单元数相关;而且,基于其间接的处理步骤(数据→网格数据→空间划分→数据划分),该方法还与数据的输入顺序无关。与基于密度的聚类只能处理数值属性的数据所不同的是,基于网格的聚类可以处理任意类型的数据,但以降低聚类的质量和准确性为代价。

STING[66]是一个基于网格多分辨率的聚类方法,它将空间划分为方形单元,不同层次的方形单元对应不同层次的分辨率。STING+ [67]则对其进行了改进以用于处理动态进化的空间数据。CLIQUE[68]也是一个基于网格的聚类算法,它结合了网格聚类与密度聚类的思想,对于处理大规模高维数据具有较好的效果。除

这些算法以外,以信号处理思想为基础的 WaveCluster[69]算法也属基于网格聚类的范畴。

（3）基于图论的聚类

基于图论的方法是把聚类转换为一个组合优化问题,并利用图论和相关的启发式算法来解决该问题。其做法一般是先构造数据集的最小生成树(Minimal Spanning Tree, MST),然后逐步删除 MST 中具有最大长度的那些边,从而形成更多的聚类。基于超图的划分和基于光谱的图划分方法[70]是这类算法的两个主要应用形式。该方法的一个优点在于它不需要进行一些相似度的计算,就能把聚类问题映射为图论中的一个组合优化问题。

（4）基于平方误差的迭代重分配聚类

基于平方误差的重分配聚类方法的主要思想是逐步对聚类结果进行优化、不断将目标数据集向各个聚类中心进行重新分配以获得最优解(判断是否是最优解的目标函数通常通过平方误差计算法得到)。此类方法又可进一步分为概率聚类算法、考虑了最近邻影响的最近邻聚类算法以及 K-medoids 算法和 K-means 算法。

1)概率聚类算法的重要代表是 Mitchell 等人于 1997 年提出的期望最大化算法(Expectation Maximization, EM) [71] 。它除了能处理异构数据之外,还具有另外几 个重要的特性: ①能处理具有复杂结构的记录; ②能够连续处理成批的数据; ③具有在线处理能力; ④产生的聚类结果易于解释。

2)最近邻距离的计算在聚类过程中起着基础性的作用,这也正是导致产生最近邻聚类算法的直接因素。共享最近邻算法(Shared Nearest Neighbor, SNN) [72]就是该类算法的典型代表之一,它把基于密度的方法与 ROCK 算法的思想结合起来,通过只保留数据点的 K 个最近邻居从而简化了相似矩阵,并且也保留了与每个数据点相连的最近邻居的个数,但是其时间复杂度也提高到了 $O(N^2)$ (N 为数据点个数) 。

3) K-medoids 方法用类中的某个点来代表该聚类,这种方法能有效处理异常数据。它的两个最早版本是 PAM 和 CLARA 算法[73],此后又有 CLARANS[74]及其一系列的扩展算法。这类方法具有两个优点:它能处理任意类型的属性;它对异常数据不敏感。

4) K-means 算法是目前为止应用最为广泛的一种聚类方法,其每个类别均用该类中所有数据的平均值(或加权平均)来表示,这个平均值即被称作聚类中心。

该方法虽然不能用于类别属性的数据,但对于数值属性的数据,它能很好地体现聚类在几何和统计学上的意义。但是,原始 K-means 算法也存在如下缺陷: ①聚类结果的好坏依赖于对初始聚类中心的选择; ②容易陷入局部最优解; ③对 K 值的选择没有准则可依循; ④对异常数据较为敏感; ⑤只能处理数值属性的数据; ⑥聚类结果可能不平衡。

为克服原始 K-means 算法存在的不足,研究者从各自不同的角度提出了一系列 K-means 的变体,如 Bradley 和 Fayyad 等人从降低聚类结果对初始聚类中心的依赖程度入手对它作了改进,同时也使该算法能适用于大规模的数据集[75]; Dhillon 等人则通过调整迭代过程中重新计算聚类中心的方法使其性能得到了提高[76] ; Zhang 等人利用权值对数据点进行软分配以调整其迭代优化过程[77] ; Pelleg 等人提出了一个新的 X-means 算法来加速其迭代过程[78] ; Sarafis 则将遗传算法应用于 K-means 的目标函数构建中,并提出了一个新的聚类算法[79]; 为了得到平衡的聚类结果,文献[80]利用图论的划分思想对 K-means 作了改进;文献[81]则将原始算法中的目标函数对应于一个各向同性的高斯混合模型; Berkhin 等人[82]将 K-means 的应用扩展到了分布式聚类。

9.2.3 基于约束的聚类算法

真实世界中的聚类问题往往是具备多种约束条件的,然而由于在处理过程中不能准确表达相应的约束条件、不能很好地利用约束知识进行推理以及不能有效利用动态的约束条件,使得这一方法无法得到广泛的推广和应用。这里的约束可以是对个体对象的约束,也可以是对聚类参数的约束,它们均来自相关领域的经验知识。该方法的一个重要应用在于对存在障碍数据的二维空间数据进行聚类。COD (Clustering with Obstructed Distance)[83]就是处理这类问题的典型算法,其主要思想是用两点之间的障碍距离取代了一般的欧氏距离来计算其间的最小距离。更多关于这一聚类算法的总结可参考文献[84]。

9.2.4 机器学习中的聚类算法

机器学习中的聚类算法是指与机器学习相关、采用了某些机器学习理论的聚类方法,它主要包括人工神经网络方法以及基于进化理论的方法。

自组织映射(Self-Organizing Map, SOM) [85]是利用人工神经网络进行聚类的较早尝试,它也是向量量化方法的典型代表之一。该方法具有两个主要特点:①它是一种递增的方法,即所有的数据点是逐一进行处理的;②它能将聚类中心点映射到一个二维的平面上,从而实现可视化。此外, 文献[86]中提出的一种基于投影自适应谐振理论的人工神经网络聚类也具有很好的性能。

在基于进化理论的聚类方法中,模拟退火的应用较为广泛, SINICC 算法[87]就是其中之一。在模拟退火中经常使用到微扰因子,其作用等同于把一个点从当前的聚类重新分配到一个随机选择的新类别中,这与 K-means 中采用的机制有些类似。遗传算法也可以用于聚类处理,它主要通过选择、交叉和变异这三种遗传算子的运算以不断优化可选方案从而得到最终的聚类结果。

利用进化理论进行聚类的缺陷在于它依赖于一些经验参数的选取,并且具有较高的计算复杂度。为了克服上述不足之处,有研究者尝试组合利用多种策略,如将遗传算法与 K-means 结合起来,并且使用变长基因编码,这样不仅能提高 K-means 算法的效率,还能运行多个 K-means 算法以确定合适的 K 值[88]。

9.2.5 用于高维数据的聚类算法

高维数据聚类是目前多媒体数据挖掘领域面临的重大挑战之一。对高维数据聚类的困难主要来源于以下两个因素:①高维属性空间中那些无关属性的出现使得数据失去了聚类趋势;②高维使数据之间的区分界限变得模糊。除了降维这一最直接的方法之外,对高维数据的聚类处理还包括子空间聚类以及联合聚类技术等。

CACTUS[89]采用了子空间聚类的思想,它基于对原始空间在二维平面上的一个投影处理。CLIQUE 也是用于数值属性数据的一个简单的子空间聚类方法,它不仅同时结合了基于密度和基于网格的聚类思想,还借鉴了 Apriori 算法,并利用 MDL(Minimum Descrip tion Length)原理选择合适的子空间。

联合聚类对数据点和它们的属性同时进行聚类。以文本为例,文献[90]中提出了文本联合聚类中一种基于双向划分图及其最小分割的代数学方法,并揭示了联合聚类与图论划分之间的关系。

9.3 聚类算法性能比较

从上面的分析介绍不难看出,这些现有的聚类算法在不同的应用领域中均表现出了不同的性能,也就是说,很少有一种算法能同时适用于若干个不同的应用背景。

总体来说,分割聚类算法的应用最为广泛,其收敛速度快,且能够扩展以用于大规模的数据集;缺点在于它倾向于识别凸形分布、大小相近、密度相近的聚类,而不能发现形状比较复杂的聚类,并且初始聚类中心的选择和噪声数据会对聚类结果产生较大的影响。层次聚类方法不仅适用于任意属性和任意形状的数据集,还可以灵活控制不同层次的聚类粒度,因此具有较强的聚类能力,但它大大延长了算法的执行时间;此外,对层次聚类算法中已经形成的聚类结构不能进行回溯处理。基于约束的聚类通常只用于处理某些特定应用领域中的特定需求。

机器学习中的人工神经网络和模拟退火等方法虽然能利用相应的启发式算法获得较高质量的聚类结果,但其计算复杂度往往较高,同时其聚类结果的好坏也依赖于对某些经验参数的选取。在针对高维数据的子空间聚类和联合聚类等算法中,虽然通过在聚类过程中选维、逐维聚类和降维从一定程度上减少了高维度带来的影响,但它们均不可避免地带来了原始数据信息的损失和相应的聚类准确性的降低,因此,寻求这类算法在聚类质量和算法时间复杂度之间的折中也是一个重要的问题。

表格 10 选取聚类算法所处理的目标数据的属性(数值型 N /类别型 C) 、算法的时间复杂度、能否处理大规模数据集、能否处理异常数据(噪声数据) 、能否处理高维数据、能否发现复杂的聚类形状、是否受初始聚类中心影响以及是否受数据输入顺序影响这八个参数,总结比较了一些有代表性的算法的性能。

表格 10 部分聚类算法性能总结与比较

聚类算法	目标数据属性	时间复杂度	能否处理大数据集	能否处理异常点	能否处理高维数据	能否发现复杂形状的聚类	是否受初始聚类中心的影响	是否受数据输入顺序的影响
CURE	N	$O(n^2_{sample})$	能	能	否	能	否	否
DBSCAN	N	$O(n\log n)$	能	能	否	能	是	是
Wave-Cluster	N 或 C	O(n)	能	能	能	能	否	否

Hyper-graphic	N 或 C	O(n)	能	能	能	否	否	否
CLARANS	N 或 C	$O(n^2)$	能	能	否	能	否	否
K-means	N	O(n)	能	否	否	否	是	是
SNN	N 或 C	$O(n^2)$	否	能	能	能	否	否
GA	N	与适应度函数相关	能	能	否	能	是	是

表格 10 中,算法的时间复杂度都是针对低维数据而言的, K-means 和 GA 也均为原始的标准算法; n 为目标数据的数目,对于 CURE 算法来讲,由于它的执行依赖于对样本集(Sample)的选择,所以其时间复杂度由样本集的数据数目来决定。

从表中反映出来的一个最突出的问题在于,这些算法绝大多数不适用于高维数据,而那些少数可以用于高维数据的算法,其时间复杂度也往往会随着维度的升高而显著增高。总之,虽然一些算法相对其他方法在某些方面的性能有了一定程度的提高,但它不可能在任何应用背景下均具有很好的结果,即几乎没有一个算法能同时在表格 10 中所示的八个方面都具有优良的性能。因此对于它们的改进还有一个相当大的空间。

9.4 分布式聚类

分布式聚类是指将分布在多个文件（或数据库）中的大型数据集中类似的样本自动聚类到多个簇，使得簇之间的样本有很大的差异性，而同一个簇内的样本比较相似。例如：某个传感器会持续产生数据，如果希望获取传感器的异常操作，可使用聚类算法将输出数据进行分类，将普通操作和异常操作会归类到不同的簇中。数据集合中各样本之间的相似度可以根据具体任务选择不同的指标，如曼哈顿距离、欧氏距离或余弦相似性等来计算两个样本之间的距离。聚类算法的基本思想通过将距离相近的项目归类到一起，可以实现对数据集的聚类。我们以 k-Means 为例，说明 Mahout 如何进行分布式 k-Means 算法。

k-Means 算法中所有的聚类对象必须表示为一组数值特性，而且用户必须指定算法聚类到簇的数量(称为 k)。每个对象被看作 n 维向量空间中的一个向量，n 表示能够用于描述聚类对象的所有特性集合。假设要将 m 个样本聚类到 k 个簇内，k-Means 算法的基本过程如下：算法初始化 k 个簇的质心点，通过多轮迭代处理更新质心位置，直到质心收敛到一个固定的点不再移动。K-means 算法是局

典型的目标函数聚类方法，它把数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。算法采用误差平方和准则函数作为聚类准则函数，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。k-Means 算法迭代的过程如图 16 所示，每次迭代包括两个步骤：首先样本找到距离最近的质心点，将其放入至该质心所在的簇；然后计算每个类别内所有样本点的平均值作为该簇的质心点。

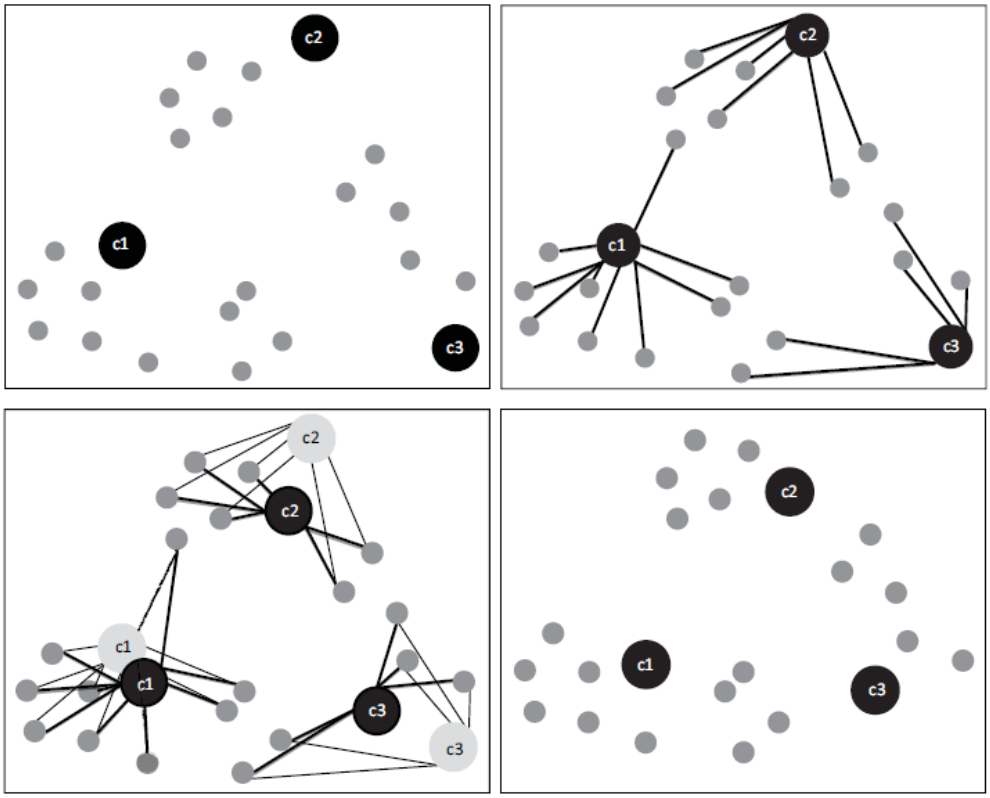


图 16 k-means 算法迭代过程

图 17 描绘了 mahout 中的 k-means 算法并行运行的过程。Mahout 提供一个 bin/ Mahout 脚本启动 k-means 聚类任务。通过设置 hadoop_home 和 hadoop_conf_dir 环境变量，可以使用相同的脚本在 hadoop 集群上启动任何 Mahout 算法。该脚本将自动读取 hadoop 集群配置文件，然后启动集群上的 Mahout 的聚类任务。具体地，整个聚类任务被当作一个 hadoop 任务来并行运行，首先每个 Mapper 程序从 SequenceFile 一个质心向量，然后每个 Mapper 计算距离每个向量最近的质心，而每个 Reducer。每个 Reducer 从每个 Mapper 获取其簇中的所有点，并且重新计算质心，然后迭代执行此过程。

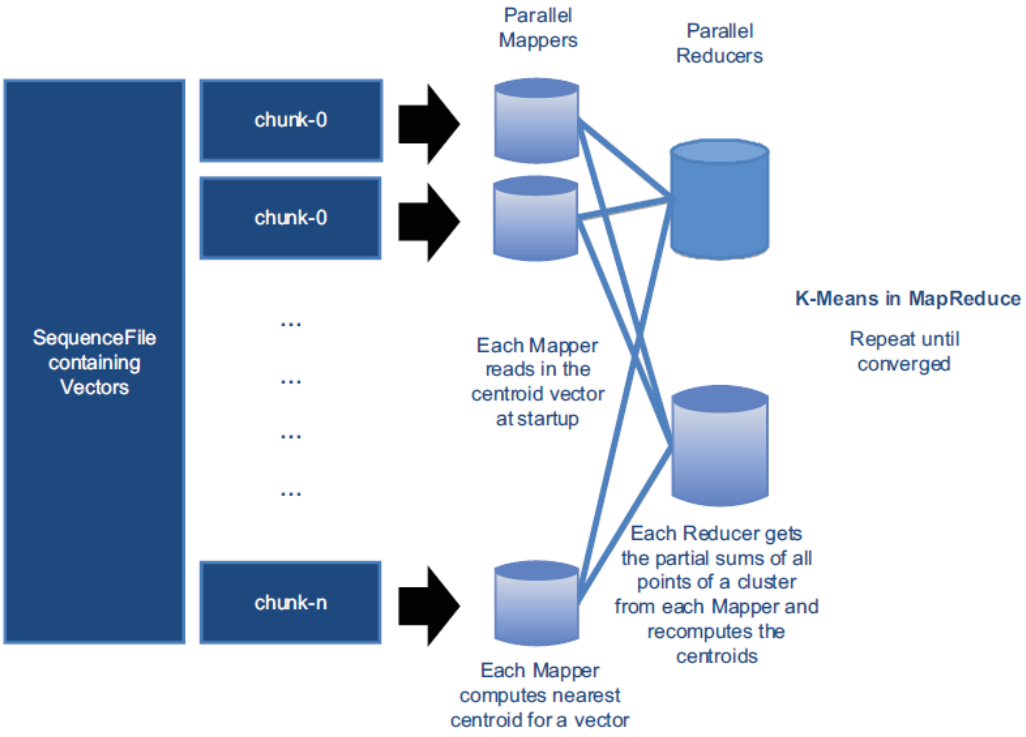


图 17 Mahout 中 k-means 聚类迭代运行于 MapReduce 平台示意图

在 Mahout 中运行 k-Means 算法，数据通常被表示为特征向量，特征向量表示了数据在 n 维空间内的一组权重值。Mahout 中的算法可以单独运行也可以部署于 Hadoop 框架下分布式进行，在分布式处理模式下使用的文件必须是 SequenceFile 格式，需要将样本数据集放到 hdfs 中指定文件下，并转换成 SequenceFile 格式才能使用 KMeansDriver 运行。SequenceFile 是 hadoop 中的一个类，允许向文件中写入二进制的键值对。图 18 是在一系列随机产生的 2 维数据点上运行 k-Means 聚类算法的结果，算法中规定簇的个数为 3，图中不同颜色的圆圈代表了每次迭代产生的簇边界，红色圆圈代表了最终结果。

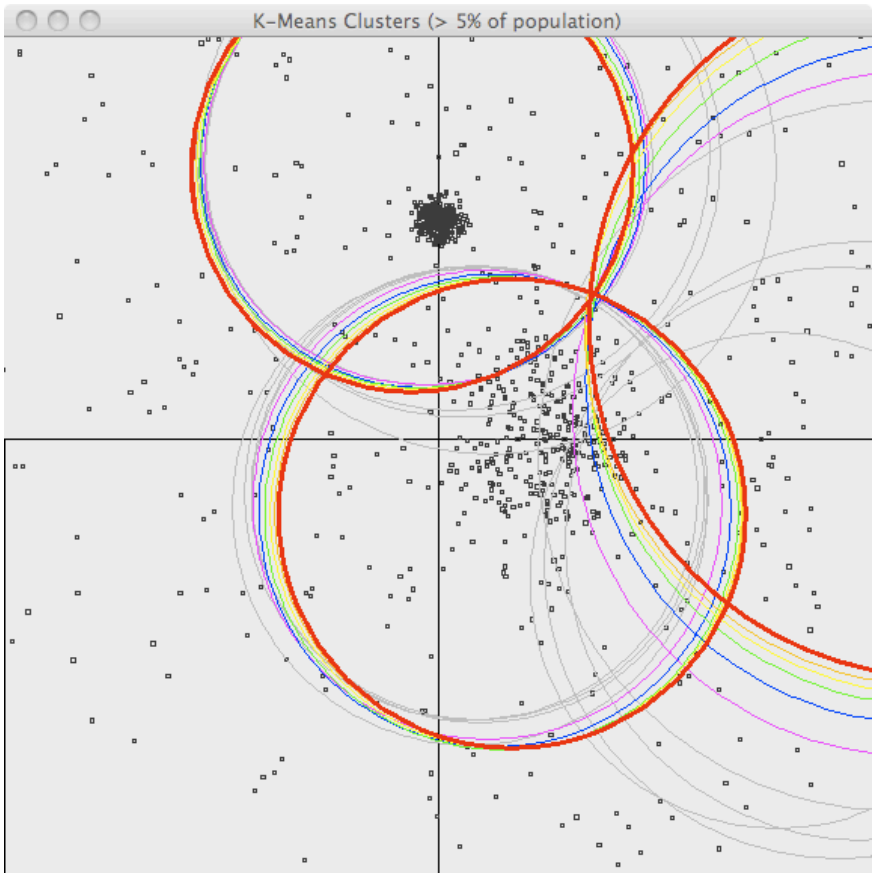


图 18 多次迭代后 k-means 聚类的结果[11]

9.5 总结

聚类算法的研究具有广泛的应用前景，其今后的发展也面临着越来越多的挑战。以其在多媒体领域中的应用为例，鉴于多媒体特征数据本身所具备的高维性、复杂性、动态性以及容易达到大规模的特性，对多媒体数据聚类算法的设计还应该更多地考虑以下几个方面的内容：

(1)融合不同的聚类思想形成新的聚类算法，从而综合利用不同聚类算法的优点。

(2)处理大规模数据和高维数据的能力，这是多媒体数据挖掘中聚类算法必须解决的关键问题。

(3)对聚类的结果进行准确评价，以判断是否达到最优解，这也自然要求聚类结果具有可解释性。

(4)选取合适的聚类类别数，这是一个重要的参数。它的确定应更多地依赖于相关的经验知识以及对目标数据集所进行的必要的预处理。

(5)对数据进行合理的预处理。该过程包括对高维数据以及对大规模数据建立索引等，它不仅是实现(4)的前提之一，也为获得更准确的聚类结果提供了一个重要的手段。

(6)在聚类过程中使用合适的相似计算公式及评价准则。合理的相似性评判准则对聚类结果的准确性起着不容忽视的作用。

(7)将领域知识引入聚类过程。领域知识的引入不仅有助于选择合适的模式表达机制、选择合适的聚类算法，还能使以上很多方面的问题都能得到合理的解决，从而提高相应的聚类算法的性能。

在多媒体数据聚类的应用中，对原始数据如图像等进行特征提取，并用这些特征数据代替原始数据进行聚类，均体现了领域知识的融合。

10 面向局域电网：节点间关联关系分析

电力系统中各种扰动引起的电能质量问题主要可分为稳态电能质量问题和暂态电能质量问题。稳态问题以波形畸变为特征，主要包括谐波、间谐波、噪声和频率波动等；暂态问题通常是以频谱和暂态持续时间为特征，可分为脉冲暂态和振荡暂态两类，主要包括电压跌落、电压骤升、短时断电和电容充电暂态等[91]。暂态问题中的电压骤降问题由于其发生的可能性远大于电压中断，即使几百公里以外的故障也有可能引起本地的电压跌落，因此在工业化国家，电压骤降已上升为最重要的电能质量问题之一，在国际上受到特别关注。国内电力企业对电压骤降的关注比其他问题电能质量问题的关注程度要高得多，同时用户对电压暂降引起问题的投诉占到全部投诉的 80%以上[92]。尤其是近 10 年来，随着高新技术、信息技术飞速发展，基于计算机控制的各种生产生活中的用电设备大量使用，对电压变化更敏感。随意电压骤降问题的日益突出，对电力系统中的电压骤降予以及时发现并判明干扰源头等工作显得尤为重要，可以为电力系统发现问题、解决问题并改进系统、提高供电质量提供工作基础[93]。

在发现问题方面，主要依据时间轴上电压、电流、频率、相位等基本监测参数变化来确定是否发生了电压暂降问题。这一工作目前主要依赖于各电子厂商生产的丰富电能质量监测仪器来完成，具有快速、准确的特点。但是，基于仪器只能判明某条线路上是否发生了电压暂降，而对于该电压暂降问题是本地线路所引起还是由相邻的其它线路所引起的问题，并不能给出有效的答案或判据[94]。

传统上对于电压事件干扰源的判定通常是基于电压事件监测数据，结合调度事件和电力系统的异常事件记录进行手工分析，具有定位精确、原因明确的优点，但同时也有耗时费力、主观性强及易受事件记录完备性影响等缺点。

当前，电能质量监测系统已经开始广泛部署，部分省网公司已经把电能质量测量延伸到 10kv 线路，未来随着用电方对电能质量问题的愈加重视以及电能质量商品化的趋势驱动，电能质量监测系统必将延伸至用户进线部分。同时，计算科学中的数据挖掘方法为分析电能质量问题带来新的技术手段。基于全网的同步电能质量监测数据，运用丰富的数据分析手段和数据挖掘方法，我们有望从数据中进行电压暂降干扰源的判断。

本文提出采用基于关联规则分析算法，面向电网局部区域化进行整体分析，从而快速定位电压暂降干扰源的判定，为高效率低成本地进行电能质量问题分析开拓一条新路径。

10.1 关联规则分析法

关联规则是挖掘数据库中两个或多个变量之间存在的隐含的关系。挖掘顾客交易数据库中项集间的关联规则问题由 Agrawal 等于 1993 年首先提出，以后人们对关联规则挖掘问题进行了大量的研究。主要研究内容是对原有算法进行优化，如引入事务压缩、杂凑、数据库划分、随机采样、并行的思想等，以提高算法挖掘规则的效率；对关联规则的应用进行推广。

10.1.1 关联规则基本定义

定义 1（关联规则）关联规则是由 Agrawal [95] 等人首先提出的一个重要 KDD 研究课题，它反映了大量数据中项目集之间有意义的关联或相关联系。

定义 2（项）设 I = { i1, i2 ,…, ik}是一个二进制数字的集合，其中的元素称为项(item)。

定义 3（支持度）记 D 为交易(transaction) T 的集合，交易 T 是项的集合，并且 T⊆I。支持度 support (A⇒B)=P(A∪B) ，其中，A⊂I，B⊂I，并且 A∩B =∅。

定义 4（置信度）confidence(A⇒B)=P(B|A)= support(A⇒ B)/support(A)，其中，A⊂I,B⊂I ，并且 A∩B =∅。

定义 5（强关联规则）是指挖掘出支持度大于客户指定的最小支持度(min_supp)和可信度大于最小可信度(min_conf)的关联规则。

定义 6（频繁项集）如果项集的出现频率大于或等于 min_supp 与 D 中事务总数的乘积，则称它为频繁项集。

10.1.2 关联规则评价及提升度

支持度是一个重要的度量，如果支持度很低，代表这个规则只是偶然出现，基本没有意义。因此，支持度通常用来删除那些无意义的规则。而置信度是通过规则进行的推理具有可靠性。对于规则 R:X⇒Y，只有置信度越高，Y 出现在包含 X 的事务中的概率才越大，否则这个规则也没有意义。

通常做关联规则会预设定最小支持度阈值 min_sup 和最小置信度阈值和 min_conf，而关联规则发现则是确定那些支持度大于等于 min_sup 并且置信度大于 min_conf 的所有规则（即“强关联规则”）。

单纯用支持度-置信度框架评价关联规则具有一定局限性。例如，如果图书市场中文学类书籍的数量远大于物理类书籍，那么物理类书籍的规则支持度就会很低，这样就导致很多物理类书籍的关联规则都被过滤掉了。再例如，如果 1000 个人中有 200 人喜欢喝茶，其中有 150 人喜欢喝咖啡，50 人不喜欢，那么我们通过置信度计算发现规则“R：喝茶⇒喝咖啡”的置信度非常高。但是可能另外不喜欢喝茶的 800 人中，有 650 人喜欢喝咖啡。由此可见喝茶和喝咖啡是两个独立事件，置信度度量忽略了规则后件中项集的支持度。

为了解决上述问题，引入了提升度(lift)的概念[]，来计算置信度和规则后件项集支持度的比率：

lift(A⇒B) = confidence(A⇒B) / support(B) =(p(A,B) / p(A)) / p(B)= p(A,B) / p(A)p(B)

lift(A⇒B)也称为兴趣因子，表示为 I(A,B)

通过概率学知识我们可以知道，如果 A 事件和 B 事件相互独立（或者我们称之为满足事件独立性假设），那么 p(A,B) = p(A) * p(B)，那么我们则可以这样来表示兴趣因子的度量：

当 I(A,B) = 1 时，我们称 A 和 B 是相互独立的，当 I(A,B) < 1 时，我们称 A 和 B 是负相关的，否则我们称 A 和 B 是正相关的。

10.1.3 关联规则的经典算法 Apriori 算法

挖掘顾客交易数据库中项集间的关联规则由 Agrawal 等于 1993 年首先提出，并设计了一个基本算法，其核心是基于频集理论的递推方法，即基于两阶段频集思想的方法，将关联规则的设计分解为两个子问题：1）. 找到满足最小支持度阈值的所有项集，我们称之为频繁项集。（例如频繁二项集，频繁三项集）；2）从频繁项集中找到满足最小置信度的所有规则。

由于步骤 2 中的操作极为简单，因此挖掘关联规则的整个性能就由步骤 1 中的操作处理所决定。挖掘关联规则的总体性能由第一步决定，第二步相对容易实现。首先产生频繁 1-项集 1，其次是频繁 2-项集 2，直到存在某个 r 值使得

频繁项集为空，此时算法停止。其中在第 k 次循环中，产生候选 k-项集的集合，中的每一个项集是由两个只有一个项不同的属于 1 的频集通过(k-2)-连接来产生的。

中的项集是用来产生频集的候选集，其中最后频集必须是的一个子集。中的每个元素需在交易数据库中需要验证来决定其是否加入，这个验证过程是影响算法性能的一个瓶颈。可能产生大量的候选集，以及可能需要重复扫描数据库，是 Apriori 算法的两大缺点。

10.1.4 FP-growth 频集算法

针对 Apriori 算法的固有缺陷，J. Han 等提出了不产生候选挖掘频繁项集的方法：FP-growth 算法[96]。采用分而治之的策略，在经过第一遍扫描之后，把数据库中的频集压缩进一棵频繁模式树（FP-tree），同时依然保留其中的关联信息，随后再将 FP-tree 分化成一些条件库，每个库和一个长度为 1 的频集相关，然后再对这些条件库分别进行挖掘。当原始数据量很大的时候，也可以结合划分的方法,使得一个 FP-tree 可以放入主存中。实验表明，FP-growth 对不同长度的规则都有很好的适应性，同时在效率上较之 Apriori 算法有巨大的提高。具体算法分为两步：

（1）构造 FP-Tree

挖掘频繁模式前首先要构造 FP-Tree，算法如下：

输入：一个交易数据库 DB 和一个最小支持度 threshold。

输出：它的 FP-tree。

步骤：

1）扫描数据库 DB 一遍，得到频繁项的集合 F 和每个频繁项的支持度，把 F 按支持度递降排序，结果记为 L。

2）创建 FP-tree 的根节点，记为 T，并且标记为’null’，然后对 DB 中的每个事务 Trans 做如下的步骤：

根据 L 中的顺序，选出并排序 Trans 中的事务项，把 Trans 中排好序的事务项列表记为[p|P]，其中 p 是第一个元素，P 是列表的剩余部分。调用函数 insert_tree([p|P],T)，其运行如下：

如果 T 有一个子节点 N，其中 N.itemName=p.itemName，则将 N 的 count 域值增加 1；否则，创建一个新节点 N，使它的 count 为 1,使它的父节点为 T，并且使它的 nodeLink 和那些具有相同 itemName 域串起来。如果 P 非空,则递归调用 insert_tree(P,N)。

（2）挖掘频繁模式

对 FP-Tree 进行挖掘，算法如下：

输入：一棵用算法一建立的树 Tree

输出：所有的频繁集

步骤：

调用 FP-growth(Tree,null)。

```
procedure FP-Growth (Tree, x)
{
    if (Tree 只包含单路径 P) then
        对路径 P 中节点的每个组合（记为 B）
        生成模式 B 并 x，支持数=B 中所有节点的最小支持度
    else 对 Tree 头上的每个 ai，do
    {
        生成模式 B= ai 并 x，支持度=ai.support;
        构造 B 的条件模式库和 B 的条件 FP 树 TreeB;
        if TreeB != 空集 then
            call FP-Growth (TreeB , B )
    }
}
```

10.2 电能质量关联规则分析建模

当前，电能质量监测仪器广泛部署，多省市已初步实现监测网络化。建有监测网的省网公司的数据中心经数年运行，已经积累了海量的电能质量问题监测数据。利用这些采集自各监测点的数据，我们有望通过比较同期电能质量问题记录，来判断不同监测节点相互之间的影响关系，从而确定电能质量干扰源所在线路或所在线路区域。

本文将基于关联分析算法定位干扰源的方法建模如下：

关联分析用于发现项集中项的关联，因而我们把每个监测点抽象为项集的一项，表现为二维数据表中的一列；

我们把所有电能质量监测节点在同一时刻上的同一种类的电能质量问题监测记录作为一个事务；

对于当前电能质量问题记录方式来讲，同一时刻同一问题的不同相位（A,B,C）的记录记为不同的事务；

对于谐波来讲，不同谐波记为不同的问题。

关联分析的结果以关联规则集合来体现，每条关联规则又体现项集之间的时间秩序关系（因果关系），在本模型中解释为节点（集）与节点（集）之间在某电能质量问题上的影响关系。

综上，本方法建模如图 1 所示。

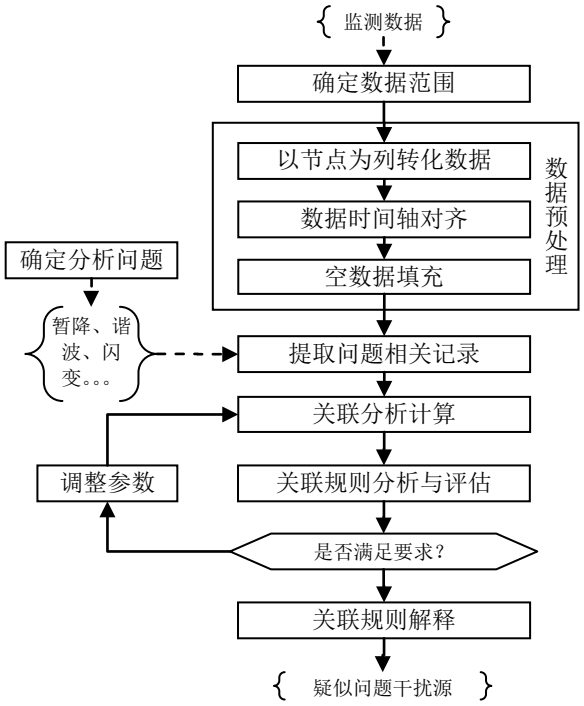


图 19 电能质量关联分析建模流程

10.3 挖掘过程的实现与分析实例

本文以江苏省电网的电能质量监测系统为基础，进行了基于关联分析的电能质量干扰源定位系统实现。

10.3.1 初始数据结构

本文实验数据来自江苏省电网监控中心的数据库，该数据采集全省 1000 多个监测点。本文用到的主要数据是电能质量历史数据超标表（dat_overrun），其基本结构如下图所示。

OID	CHV_OID	CHI_OID	DAT_OID	KIND	OC_DATE	SEQ	COL
1304382424	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"11"
1304382425	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"24"
1304382426	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"25"
1304382427	0	3131	6898101203	6	"2012-08-15 00:20:00"	"A"	"24"
1304382428	0	3131	6898101203	6	"2012-08-15 00:20:00"	"A"	"25"
1304382429	0	3131	6898101204	6	"2012-08-15 00:30:00"	"A"	"24"
1304382430	0	3131	6898101204	6	"2012-08-15 00:30:00"	"A"	"25"
1304382431	0	3131	6898101205	6	"2012-08-15 00:40:00"	"A"	"24"
1304382432	0	3131	6898101205	6	"2012-08-15 00:40:00"	"A"	"25"
1304382433	0	3131	6898101206	6	"2012-08-15 00:50:00"	"A"	"24"
1304382434	0	3131	6898101206	6	"2012-08-15 00:50:00"	"A"	"25"

图 20 电能质量历史数据超标表结构示意图

其中，chi_oid 是线路监测点 ID，kind 代表电能质量问题类型，其中

1 代表：频率，

2 代表：短闪变，

3 代表：长闪变，

4 代表：电压不平衡，

5 代表：谐波电压，

6 代表：谐波电流，

7 代表：总谐波畸变率，

8 代表：电压变动，

9 代表：电压偏差’

oc_date 是记录时间，seq 代表出问题的相位，col 字段用于记录谐波波次。总计记录数据为 4232007 条。

另有电压通道表 DV_CHV、母线关联 DV_CHV_REF、电流通道表 DV_CHI、线路关联 DV_CHI_REF、变电站(监测网) DV_STATION、变电所关联 DV_STATION_REF 等系统级数据表用于记录各监测点在电网中的位置信息。

10.3.2 数据预处理

本文面向整个省网监测系统，因而首先从原始数据集中提取包含的所有监测节点（以下简称“节点”），以此生成新数据集的列。

原始数据集中，除谐波电流问题记录用 chi_oid 唯一标识一个节点外，其余电能质量问题均用 chv_oid 唯一标识一个节点。本文没有拿到进一步资料说明

chi_oid 与 chv_oid 是一一对应并同标识的，因而，本文将数据分成谐波电流和其它问题两个部分处理。新表数据结构如下：

表格 11 预处理数据表结构

字段	类型	值域	说明
Time	datetime		时间点
Kind	Int	1-9	问题类型
Seq	Char	T,A,B,C	相位
Col	Int	2-50	谐波波次
Node_7	tinyint	0,1	节点 7 发生问题否
Node_8	tinyint	0,1	节点 8 发生问题否
…	tinyint	0,1	其它节点

生成两个数据表，其中谐波电流表（记为 chi_overrun）337 个字段，包含 333 个节点（项），其它问题表（记为 chv_overrun）544 个字段，包括 540 个节点。本文以 python 编程方式生成这两个表。

对于来集自监测系统的原始数据，通过反复扫描数据表，将同一问题同一时间点上所有节点的数据写入新数据表的同一条记录，未发生相应问题的监测节点对应的列填缺省数值 0。原始数据经数据转换后填入新表的情况如表 2 所示。

表格 12 预处理后数据记录变化对比

问题类别	原始记录数	转换后记录数
1 频率	846	179
2 短闪变	13380	3720
3 长闪变	12935	3793
4 电压不平衡	2666	1422
5 谐波电压	2653366	227463
6 谐波电流	1255229	197847
7 总谐波畸变率	57858	6500
8 电压变动	140975	1772
9 电压偏差	94752	6491

10.3.3 基于 FP-growth 算法进行挖掘

从转换预处理后的数据集 chv_overrun 中，提取关于电压变动的所有数据，去掉时间（Time)、问题（Kind)、相位（Seq）、波次（Col）列，存为 CSV 文件。

在 weka3.7 的 Explorer 组件中打开该文件，使用非监督学习中的属性过滤器 NumericToNominal 对数据值型数据（0，1）进行离散化处理。之后，我们选择 Associator 中的 FP-growth 算法进行节点关联性挖掘。

其中 FP-growth 涉及的参数包括如下：

表格 13 FP-growth 算法参数说明表

参数	含义	本文设置
Delta	每次迭代中最小支持度的增值幅度	5
findAllRulesForSupportLevel	是否要找出满足支持度的全部规则	否
lowerBoundMinSupport	支持度的最小值	5
maxNumberOfItems	项集的最大项数	2
metricType	是指定选择哪个量进行排序,weka 提供四种排序方法，0=confidence ， 1=lift ， 2=leverage ， 3=Conviction；	0
minMetric	指你选定的那个排序参数的那个最小值	0.5
numRulesToFind	给出要输出多少条规则	20
positiveIndex	正值的属性索引，对于密集数据的二元属性索引设为“正”，对于稀疏数据属性索引总是设为“2”	
upperBoundMinSupport	支持度的最大值	80

由于相邻线路才会产生直接影响，监测系统中提供了各监测节点所在线路信息。根据各线路所属的变电站，我们可以重点计算变电站内各线路之间的影响，再计算相邻变电站之间的线路影响。这样可以极大地降低计算复杂度。基于这样的原则，我们选择了相邻近的 50 个监测节点的数据进行分析。本文所示例的计算问题为电压偏差。

10.3.4 规则解释

经挖掘后，共产生 387 条关联规则（如下图所示）。

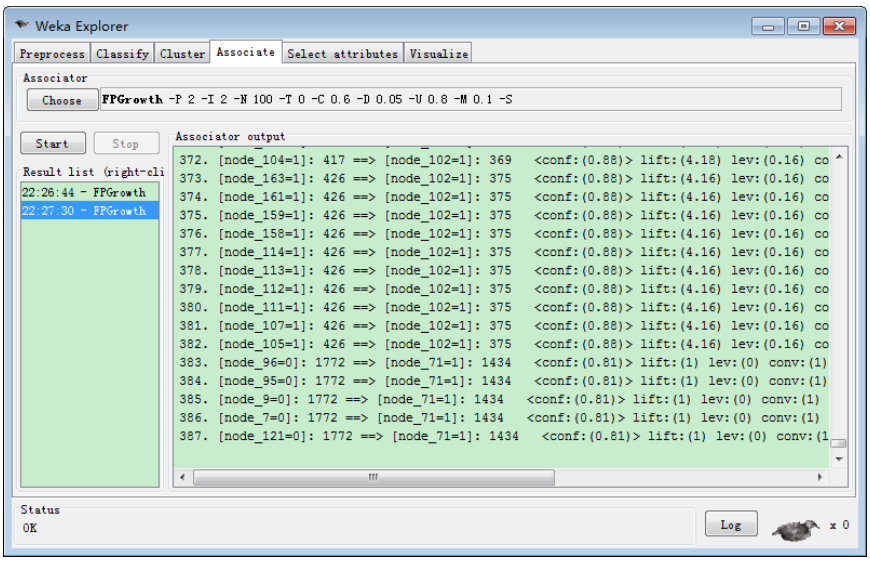


图 21 FP-growth 挖掘结果示意图

去除（0==>0, 0==>1, 1==>0）这样的非兴趣规则 115 条，通过对剩余 272 条规则进行分析，我们得出以下 2 条规律：

1）节点编号为 102、103、104、105、106、107、111、112、113、114、160、162、164 的 13 个节点在电压偏差问题具有较明显的共现关系，因而在其中一点发生电压偏差问题时，应向其它节点发出警报，或自动启动其它节点的防范动作；

2）节点编号为 103、104、105、106、107、111、112、113、114、158、159、160、161、162、163、164 等 16 个节点的电压偏差问题对节点 102 有明显的影响作用。

10.4 基于时间序列关系的谐波源定位方法

谐波对谐波使电能的生产、传输和利用的效率降低，使电气设备过热、产生振动和噪声，并使绝缘老化，使用寿命缩短，甚至发生故障或烧毁。谐波可引起电力系统局部并联谐振或串联谐振，使谐波含量放大，造成电容器等设备烧毁。谐波还会引起继电保护和自动装置误动作，使电能计量出现混乱。对于电力系统外部，谐波对通信设备和电子设备会产生严重干扰[97]。为了及时解决配电网中的谐波污染问题，达到分清谐波责任，简单有效的治理目的，正确识别综合负荷中的主要谐波源是至关重要的。

传统上，谐波源定位是通过测量某些点（如公共连接点 PCC）的电压、电流或功率值，在所测数据的基础上，采用相应的算法判定系统侧和用户侧谁是主要谐波源。若系统侧为主要谐波源，则对电压、电流畸变负主要责任；反之，则用户侧应承担主要责任。基于功率方向的方法简单直观、易于实现。然而，有功功率方向法[98]易受 PCC 两侧电压相角差的影响，不能正确判断主谐波源位置。无功功率方向法[99]和临界阻抗法[100]等方法易受谐波阻抗估计值的影响；基于谐波阻抗的方法[101,102,103,104,105,106]原理简单、清晰。然而，它的前提难以实现，因为谐波阻抗是在扰动情况下测量的，实际中的扰动具有随机性，很不稳定。

当前，专用电能质量监测仪器已经被广泛部署于电网之中。经多年运行，电能质量监测网已经积累海量的监测数据。其中关于谐波的监测数据比较全面地反

应了各个时段各个波次的谐波分布情况。在监测网点分布全面合理的情况，具有了利用监测数据来挖掘分析谐波源的定位实现的可行性。

本文提出以网络拓扑分析和叠层式序列比较法来进行谐波源定位的方法。在不对电网进入任何额外的电流注入和电压干扰的情况，利用已有的数据，可以快速的定位出谐波源的来源范围，为进一步精确定位谐波源极大地消减了工作量。由于监测的持续性和实时性，本方法对于新出现的谐波源能够及时发现。本方法来用了来自江苏省电网的电能质量监测数据进行实现验证，结果证明多数情况下，计算范围能够快速收敛，准确定位到 1 点或几点的小范围之内。

10.4.1 谐波源基本性质及其建模

谐波产生的根本原因是系统中某些设备和负荷的非线性特性，即所加电压与产生的电流不呈线性关系而造成波形畸变。理想的公用电网所提供的电压应该具有单一而固定的频率和规定的电压幅值，但当系统的正弦波形电压加在非线性负载上时，产生的电流为非正弦波形，波形的畸变即产生了谐波电流，谐波电流又影响端电压，使电压波形发生畸变从而产生谐波电压。这些向电网中注入谐波电流和产生谐波电压的电气设备即为谐波源[107]。

电网中的谐波源大体分为两种类型：一类为含有半导体元件的各种电力电子设备，如各种整流、逆变装置和晶闸管可控开关设备等，它们按一定的规律开闭不同电路，将谐波电流注入电网；另一类为含有电弧和铁磁非线性设备的谐波源，如荧光灯、电弧炉和各种铁心设备包括变压器、电抗器等。家用电器设备分属于上述两类谐波源，虽然其容量小，但数量巨大，因此也是不可忽视的谐波源。此外，对于电力系统三相供电来说，三相不平衡负荷也是典型的谐波源，使电力系统的电流和电压波形产生畸变。

对于各种非线性负荷的谐波建模，IEEE 谐波工作组报告指出，谐波源的特性可以表述为 $I_h = F_h(V_1, V_2, V_3, \ldots)$，并且推荐了各 $h = 1, 2, \ldots, N$，种非线性负荷的谐波源模型用于谐波分析[108]。在电压波形畸变严重或电压不平衡的情况下，则需要更详细的谐波负荷模型[109]。这些模型虽然比较精确，但是在建模过程中需要各种负荷的详细参数。

在实际情况下，由于负荷种类繁多，很难精确获得各种所需参数，而这些模型的运算又比较复杂，因此在研究中常采用简化模型，包括：

(1) 恒流源模型[110]；

(2) Norton 等效电路模型[111]；

(3) 基于交叉频率导纳矩阵的简化模型[112]；

(4) 基于电压基波零相角特性的谐波源简化模型[113]。

这些简化模型各具特点，同时在实际的研究应用中也存在不同的限制。此外，不断地有新的理论和方法如瞬时功率理论、统一参数辨识方法等[114,115]运用到负荷的谐波建模中。广义谐波理论下的负荷建模也被尝试应用于负载线性程度的研究[116]。

10.4.2 传统谐波源辨识方法

谐波源的辨识问题最初是作为谐波潮流的逆问题由 G. T. Hey dt 提出[117]的。通过测量系统中部分节点的谐波电压和线路中的谐波电流，采用状态估计的方法来获得负荷注入系统的谐波功率。当注入的谐波功率为正时，则判定该负荷为谐波源。这就是目前实际应用最广泛的功率方向法。此后，很多新的算法如线性神经元网络等被运用于谐波状态估计[118]，同时量测量及状态变量的选择也各有不同，并在此基础上开发了相应的实用程序和装置用于实际辨识电网中的谐波源[119]。此外，瞬时功率理论也被尝试运用于谐波源辨识研究，通过三相系统中特定节点处滤波装置数据得到系统瞬时谐波有功功率，来定位电网中主要的谐波源。日本学者据此研制出相应的装置并用于试验研究[120]。但是对于单相网络中存在的谐波源或因单相负荷所引起的谐波畸变，它不能提供任何有用的信息。而基于瞬时无功功率理论的辨识方法在这方面有所改善[121]。这些研究主要集中在状态估计方法上，旨在利用最少的测量装置取得令人满意的识别结果，对于谐波源识别的判据并没有新的观点。

另一类辨识谐波源的方法是通过研究畸变的电压波形和电流波形之间的内在联系，找出相应的负荷参数，作为判定谐波源的指标，如基于外加负荷扰动法和基于瞬时负荷参数分割法。前者根据负荷在外加扰动的情况下，其谐波电流、基波电流和谐波电压三者幅值之间的相互关系来判断负荷中是否含有谐波源[122]。如果按照这一关系而在相应的坐标系中绘制的点图中，谐波电流和同次谐波电压的关系可以拟合为过原点的一条直线，而谐波电流和基波电流无关，则将负荷视为线性负荷，反之则认为负荷中存在谐波源。后者根据计算 RL 参数，

并引入非线性瞬时指标用于计算非线性统计因子 NHL 来进一步判定是否为谐波源[123] 。

10.4.3 基于监测数据的谐波源定位模型

当前，电能质量监测仪器广泛部署，多省市已初步实现监测网络化。建有监测网的省网公司的数据中心经数年运行，已经积累了海量的电能质量问题监测数据，其中包含有线路上的谐波电流记录、母线上的谐波电压记录以及总谐波畸变率记录。利用这些采集自各监测点的数据，我们有望通过比较同期谐波分布状态，来判断谐波源具体位置或所在范围。

基于监测数据反应的是谐波发生的结果这一事实，我们有以下反应谐波传递规律的基本理论假设：

1）每个监测点可以抽象为网络中的一个节点；

2）每个节点只可能受相邻节点的影响；

3）若两个相邻节点在谐波的发生时间上有明显的顺序关系，则可判定处于顺序关系中较早时间的节点是较晚时间的谐波上游。

4）两个相邻节点可能在不同时间段内形成不同的上下游关系，那么这两种关系同时成立。

基于上述假设本文提出了新谐波源定位模型，模型分为两个部分：（1）计算上下游模型和（2）计算源头模型。

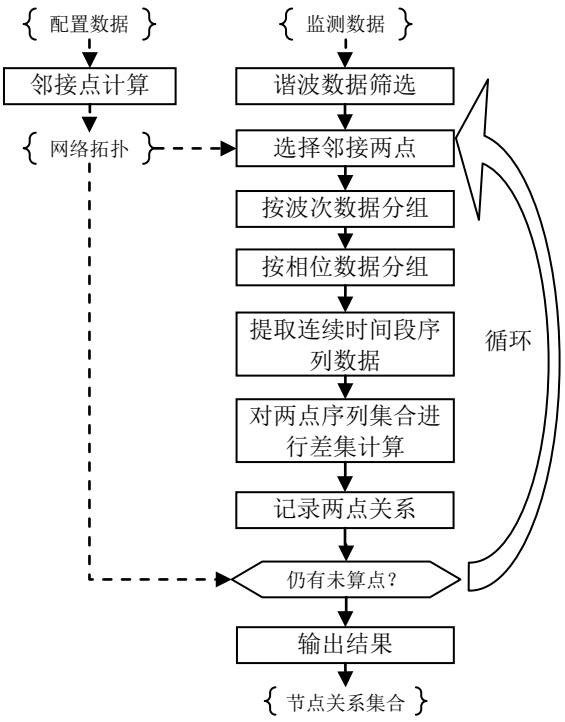


图 22 监测节点谐波关系计算流程

模型 1 用于计算监测网中各节点之间的在不同时间段内的谐波上下游关系，模型如图 22 所示。其主要计算工作在于比较任意两相邻节点的谐波间隔序列集合之间的包含关系，输出结果用于模型 2 的计算。

模型 2 用于计算某特定时间点的谐波源头，其基本思想是以模型 1 计算出的谐波有向网络集合为基础，在某个时间点的切片上计算求出网络中每个入度为 0 的节点。其基本过程如图 23 所示。

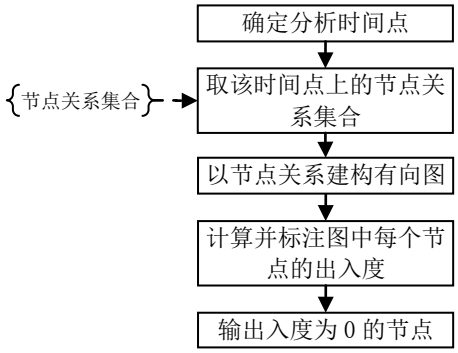


图 23 谐波源定位模型

10.4.4 谐波源定位的系统实现

本文以江苏省电网的电能质量监测系统为基础，进行了基于数据的谐波源定位系统实现。

（1）初始数据结构

本文实验数据来自江苏省电网监控中心的数据库，该数据采集各省 1000 多个监测点。本文用到的主要数据是电能质量历史数据超标表（dat_overrun），其基本结构如图 24 所示。

OID	CHV_OID	CHI_OID	DAT_OID	KIND	OC_DATE	SEQ	COL
1304382424	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"11"
1304382425	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"24"
1304382426	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"25"
1304382427	0	3131	6898101203	6	"2012-08-15 00:20:00"	"A"	"24"
1304382428	0	3131	6898101203	6	"2012-08-15 00:20:00"	"A"	"25"
1304382429	0	3131	6898101204	6	"2012-08-15 00:30:00"	"A"	"24"
1304382430	0	3131	6898101204	6	"2012-08-15 00:30:00"	"A"	"25"
1304382431	0	3131	6898101205	6	"2012-08-15 00:40:00"	"A"	"24"
1304382432	0	3131	6898101205	6	"2012-08-15 00:40:00"	"A"	"25"
1304382433	0	3131	6898101206	6	"2012-08-15 00:50:00"	"A"	"24"
1304382434	0	3131	6898101206	6	"2012-08-15 00:50:00"	"A"	"25"

图 24 电能质量历史数据超标表结构示意图

其中，chi_oid 是线路监测点 ID，kind 代表电能质量问题类型（其中，5 代表谐波电压，6 代表谐波电流，7 代表总谐波畸变率），oc_date 是记录时间，seq 代表出问题的相位，col 字段用于记录谐波波次。

另有电压通道表 DV_CHV、母线关联 DV_CHV_REF、电流通道表 DV_CHI、线路关联 DV_CHI_REF、变电站(监测网) DV_STATION、变电所关联 DV_STATION_REF 等系统级数据表用于记录各监测点在电网中的位置信息。通过这些数据表可以得出各监测节点构成的拓扑网络（本文不在此细述生成方法）。

（2）单节点谐波序列生成

每个节点记录的谐波数据可按波次、相位及发生时间分解成一系列的时间间隔连续的序列。原始数据集中，是按每 3 分钟一个统计周期来存储谐波数据的，因而，当指定监测节点、波次（如 3 次谐波）和相位的情况下，如果我们发现有两条记录的时间点为间隔 10 分钟，则认为它们是在同一时间序列中。

我们实现了一个带有游标的序列转化程序，通过在数据库中反复遍历列表，获得各个点的谐波序列集合。为减少遍历计算量，我们把初始数据备份到 MySQL 数据库当中，针对备份数据进行操作。对于遍历过的记录直接删除，只保留其序列版本。其基本处理过程如下图所示（该过程以谐波电流分析为例）。

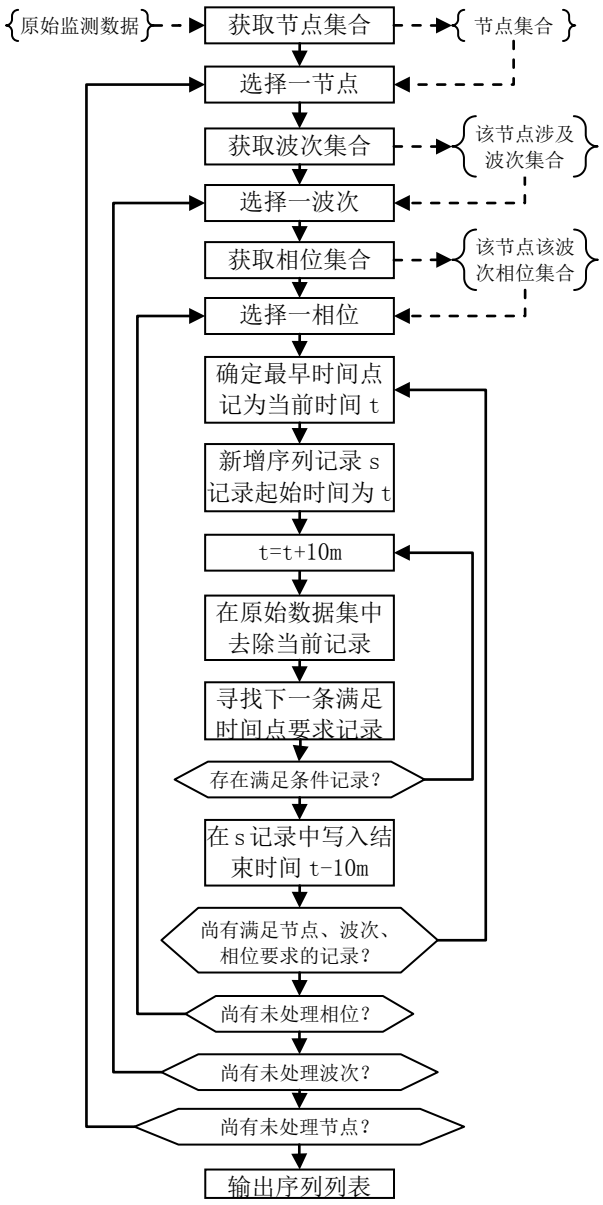


图 25 单节点谐波序列生成流程图

输出后的谐波序列如下表所示。

oid	kind	col	phase	start_time	end_time
3131	6	23	A	2012/8/15 3:40:00	2012/8/15 6:50:00
3131	6	24	A	2012/8/15 0:10:00	2012/8/15 6:50:00
3131	6	24	B	2012/8/15 0:10:00	2012/8/15 6:50:00
3131	6	24	C	2012/8/15 0:10:00	2012/8/15 7:00:00

图 26 谐波序列存储示意图

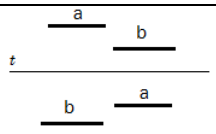
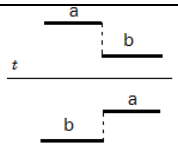
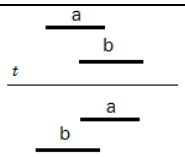
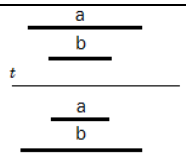
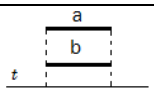
其中，start_time 代表序列起始时间，end_time 代表序列结束时间，其它字段定义与 5.1 节相同。

（3）谐波序列关系比较

由于各谐波源的特征不同，可产生不同波次、不同相位的谐波，甚至在同一时间段内产生多种不同波次的谐波，因而，两个节点的谐波序列比较一定要在相同波次、相同相位的情况下才有意义。

在同波次、同相位的情况下，本文如下定义，设有两谐波序列 a(ts1,te1)，b(ts2,te2)，ts1,te1,ts2,te2 分别为两个序列起始时间和结束时间，则 a，b 可能存在如下关系：

表格 14 谐波序列关系定义表

名称	条件公式	图例
无关	te1 <ts2 te2 <ts1	
接续	b 接续 a: te1 =ts2 a 接续 b: ts1 =te2	
交叠	a 交叠 b: ts1 <ts2 and te1 >ts2 and te1 <te2 b 交叠 a: ts1 >ts2 and ts1 <te2 and te1 >te2	
包含	a ⊃b: (ts1≤ t s2 and te1>te2) or (ts1 < t s2 and te1 ≥te2) b⊃a: (ts1 ≥ts2 and te1 <te2) or (ts1 > t s2 and te1 ≤ te2)	
相等	ts1 = t s2 and te1 = te2	

由上表中的条件公式，针对图所示的谐波序列存储表，可以容易得计算出任意两个序列的关系。

（4）节点上下游关系计算

对于相邻的两个节点 a 和 b，对其确定某次谐波的某个相位上的谐波序列集合进行比较，假设 a 点的序列集合为 A，b 点的序列集合为 B，若：

1）∀s1∈A 和 s2∈B，若 s1 与 s2 无关，则称 a 点与 b 点无关；

2）若∀ s1∈A，∃s2∈B，使得 s1 包含 s2,或 s1 交叠 s2,则称 a 点是 b 点的谐波上游，反之亦然；

3）若∀ s1∈A，∃s2∈B，使得 s1 与 s2 相等，则称 a 点与 b 点谐波等效；

4）若∀ s1∈A，∃s2∈B，使得 s2 接续 s1，则称 b 点为 a 点的疑似谐波下游,反之亦然。

除去以上4种绝对情况，在实际计算中可在去除无关序列后再统计情况2，3，4的出现和所占百分比情况，两节点的关系取决于统计周期内占百分比较多的序列关系。

（5）谐波源头计算

将（4）计算所得各节点关系，转化为图中有向边，则原始数据集被转化为一个有向图。转化规则为：1）从谐波上游节点到下游节点构建一条有向边；2）若两节点谐波等效，则在两节点之间构建一条双向边。

构建后的有向图可能由一个连通的图或几个不连通的几个子图组成（如下图所示）。其中入度为0的节点即为疑似谐波源（图中虚线圆所圈出节点）。

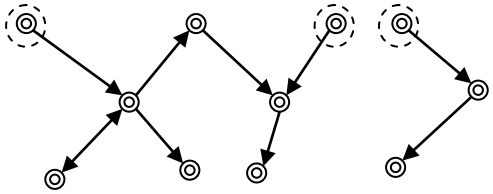


图 27 单节点谐波序列生成流程图

（6）实验验证

为验证上述基于时间序列数据分析定位谐波源方法的有效性，本文于某省电网下辖110kv变电站采集5个点的监测数据，进行了谐波源定位分析。该5点拓扑结构如下图所示。

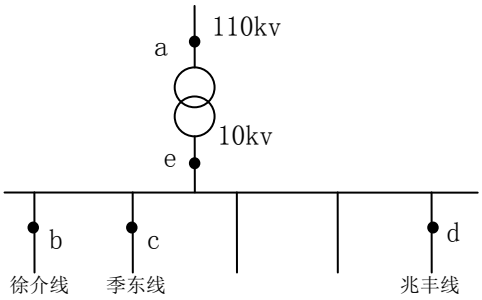


图 28 实验测试环境部署图

a点：100kv进线

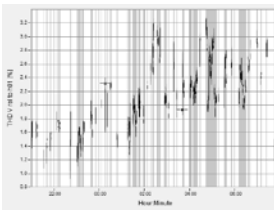
b点：居民用电为主的线路

c点：居民用电为主的线路

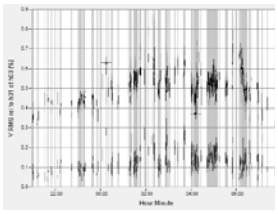
d点：已知挂有炼钢企业线路，且测试期间炼钢企业按计划生产

e点：10kv出线。

本文测取了自晚 8 点到早 8 点共计 12 小时的谐波监测数据，其中 d 点与 e 点含有谐波含量数据示例如下图。



d 点谐波含量



e 点谐波含量

图 29 谐波含量测试示例

由于a点没有监测到谐波数据，因而可以认为100kv进线没有受到谐波影响，故而在进一步的分析中放弃对 a 点数据的比较。我们以 6 次谐波电流为例，经测量及序列提取计算，得出如下表所示谐波序列表：

表格 15 各节点提取序列

节点	谐波记录	A 序列	B 序列	C 序列
b	168	29	0	32
c	539	46	0	47
d	958	19	0	20
e	753	36	0	38

按 A,B,C 三相位分别比较各节点序列关系，得到如下表所示。行节点 A 与列节点 B 的“i/j”关系中，i 代表节点发生影响关系的次数（包括 A 交叠 B 和 A 包含 B），j 代表发生被影响关系的次数（包括 B 交叠 B 和 B 包含 A）。

表格 16 A 相各节点序列关系表

	B	c	d	e
b		7/11	0/28	0/14
c	11/7		10/38	9/24
d	28/0	38/10		23/13
e	14/0	24/9	13/23	

表格 17 C 相各节点序列关系表

	b	c	d	e
b		7/9	0/31	3/15
c	9/7		10/36	10/22
d	31/0	36/10		24/13
e	15/3	22/10	13/24	

累加三相序列关系记录，并计算影响与被影响关系的比例，得到下表：

表格 18 节点序列关系统计表

	b	c	d	e
b		0.70	0.00	0.10
c	1.43		0.27	0.41

d	(=59/0)	3.70		1.81
e	9.67	2.42	0.55	

从上表我们可以看到：d 点对 b、c 点有明确的影响作用；e 点对 b 点、c 点也有明确的影响作用，但影响作用比 d 点对 b 点、c 点的影响作用略弱；d 点和 e 点相互影响，但 d 点对 e 点的影响作用更强一些。b，c 两节点间的影响较弱，不分主次。

经上述计算对比，d 为疑似 6 次谐波源。这一结论与事先所知 d 点监测线路挂有谐波源事实相符，因而证明本文方法在识别谐波源上具有一定有效性。

10.4.5 结论

本文通过对历史监测数据的统计分析，能够快速有效地定位出谐波源的所在范围，同时该方法并未对实际运行电网产生任何额外干扰。

本文只在小范围简单条件下进行了实验验证，对于更大范围多谐波源并存的情况，则需要对本文进一步精确和完善。

随着电能质量监测网络的深入发展，电能质量监测仪必然更加广泛深入地部署到具体负荷和供电设备上。届时，将本文方法在更深入范围的监测数据上计算应用，将可精确确定产生谐波的具体线路或设备。

10.5 小结

本文通过对阶段性的历史监测数据的统计分析，能够快速有效地定位出谐波源的所在范围，同时该方法并未对实际运行电网产生任何额外干扰。

随着电能质量监测网络的深入发展，电能质量监测仪必然更加广泛深入地部署到具体负荷和供电设备上。届时，将本文方法在更深入范围的监测数据上计算应用，将可精确确定产生电能干扰源的具体线路或设备。

11 面向交互分析：数据可视化技术

11.1 概述

视觉信息是人类最主要的信息来源。研究结果表明：人类日常生活中所接受的信息约 80%来自于视觉信息。基于人类强大的视觉潜能，二十世纪八十年代后期，一项新的技术——可视化技术（Visualization Technology）被提出并随后得到迅速发展。可视化是利用计算机图形学和图像处理技术，将数据转换成图形或图像在屏幕上显示出来并进行交互处理的理论、方法和技术。其涉及到计算机图形学、图像处理、计算机视觉、计算机辅助设计等多个领域，成为研究数据表示、数据处理、决策分析等一系列问题的综合技术。

目前，可视化技术主要包括科学计算可视化和数据可视化两个分支：

(1) 科学计算可视化

科学计算可视化（Visualization in Scientific Computing）是 1987 年被发达国家提出后逐渐发展起来的一个新的研究领域。通过科学计算可视化来启发和促进对自然规律的更深层认识，从而发现规律并应用于生产领域。科学计算可视化是指应用计算机图形学和图像处理技术，将在科学计算过程中产生的数据和计算结果转换为图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术[124]。实际上，随着技术的不断进步，科学计算可视化的含义已经大大扩展，不仅包括科学计算数据的可视化，还包括工程计算数据的可视化以及测量数据的可视化等。科学计算可视化是覆盖多门学科的研究领域，融合了计算机图形学、图像处理学、科学与符号计算、计算机视觉等领域的知识。

(2) 数据可视化

数据可视化（Data Visualization）是关于数据之视觉表现形式的研究。其中，这种数据的视觉表现形式被定义为一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元元素表示，大量的数据集构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察数据，从而对数据进行更深入的观察和分析[125]。

数据可视化技术包含以下几个基本概念[126]：

(1) 数据空间 (Data Space)：也称作多维数据空间，是由多维属性和多个元

素组成的数据集构成的多维空间。

(2) 映射空间 (Mapping Space)：也称作投影空间，是将多维数据按照一定的函数或规则转换后得到的低维可视空间。

(3) 多维数据分析 (Multidimensional Data Analysis)：是指对多维数据进行切片、切块、旋转等动作剖析数据，从而能多角度多侧面地观察数据。

(4) 多维数据探索 (Multidimensional Data Exploration)：是指利用一定的算法和工具对多维数据蕴涵的信息进行搜索，得到有用、新颖的信息。

(5) 多维数据可视化 (Multidimensional Data Visualization)：是指将大型数据集中的数据以图形或图像形式表示，并利用数据分析和挖掘工具开发其中未知信息的处理过程。

数据可视化技术主要有以下三个特点[127]：

(1) 交互性：用户可以方便地以交互的方式管理和处理数据。

(2) 多维性：应用对象或事件的数据具有多维属性和变量，而数据可以按其每一维的值对其进行分类、排序、组合和显示。

(3) 可视性：数据可以通过图象、曲线、二维图形、三维体和动画等形式来显示，并可对其模式和相互关系进行可视化分析。

在实现数据可视化之前，通常要求先对数据空间中的数据进行预处理，这些处理技术包括[128]：(1) 对定性数据的处理；(2) 数据的权重；(3) 数据的无量纲化处理；(4) 标准化处理。

(3)可视化技术的分类

数据的可视化将涉及到数据类型、可视化技术以及对数据进行交互和变形的技术[129]，这三个要素构成了对数据的可视化。图 30 描述了三个要素各自所包含的内容：

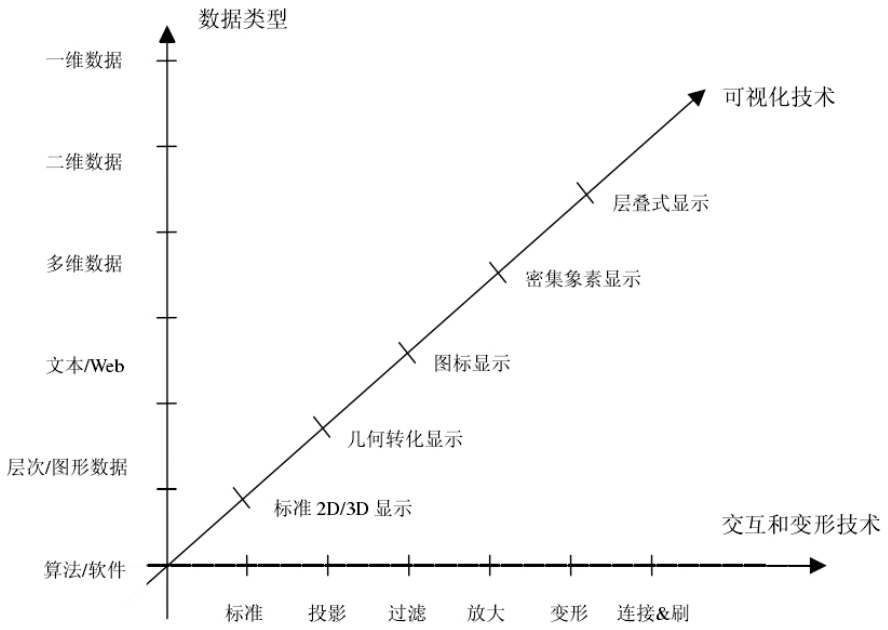


图 30 数据可视化的三要素 [130]

11.2 可视化数据类型

可视化数据的类型主要有：

(1) 一维数据：一维数据通常含有一个数据维，一般可以用一维坐标轴上的点表示，也可用二维坐标平面上的点或柱状图等表示[131]。典型的一维数据有时序数据等；

(2) 二维数据：二维数据具有两个不同的数据维，例如 X-Y 坐标；

(3) 三维数据：很多数据集由多于两个属性组成，三维数据可以用二维空间的透视图或投影图来表示[132]。有时，多维数据包含的属性多达几十到几百个，通常不能简单地运用二维或三维散点图等来可视化，一般需要运用可视化技术将多个属性映射到二维或三维空间，然后才能在屏幕上进行可视化表示。典型的三维数据如数据库中的表；

(4) 文本和超文本：这种类型的数据不能被轻易地描述为数字，因此许多标准的可视化技术不能被应用。多数情况下，首先把该类型数据转化为向量描述，然后再应用可视化技术。

(5) 其他数据类型：例如图形、层次数据、算法和软件等等。图形可以表示一般数据之间的内部依赖关系。层次数据类型可视化的相关内容可参见文献[131]。算法和软件可视化的目的是为了帮助对算法的理解，以此来支持软件的开发，例如流程图、代码结构图等等。

11.3 数据可视化技术

人们所熟悉的传统的数据可视化技术包括：折线图、柱状图、条形图、散点图、饼图、分位数图、回归曲线图等。而当前国际上流行的数据可视化技术，根据其构建和显示原理可以划分为几何显示技术、像素显示技术、图标显示技术和层次显示技术[133]。

11.3.1 传统的数据可视化技术

传统的可视化技术主要有：

(1) 折线图 (Line Graph/Chart)

折线图可表示评估测试的提升图，用于比较各种分类算法。其最简单的实现方法为：首先在 X-Y 坐标系中描出数据点，然后尽可能用线段将所描绘的点连接起来。X 轴的数据值可以是离散的，也可以是连续的。如果数据是离散的，离散值就成为X轴上依次排列的位置标签（Label），Y轴的数据值则必须是连续的。折线图通常可用于显示一个字段（数据维）的值与另一个字段（数据维）的值在 X-Y 坐标系中的对比情况或者用来描绘时间序列上的趋势。折线图是一种常用的图形，也是一种基本图形，其有很多的变种，例如雷达图等，为数据的显示带来了极大的便利。

图 31 折线图 [134]

(2) 柱状图和条形图 (Column and Bar Graphs)

这两种图都可用于在 X-Y 坐标系中比较离散数据维和连续数据维交叉点的值。其中，柱状图绘制数据的方式类似于折线图，均是在离散字段和连续字段的交叉点处画出数据点。其和折线图的区别在于：前者比后者多了在 X 轴上的垂直圆柱（Column）来表示数据维的值。条形图和柱状图的本质相同，只是两者的 X 轴和 Y 轴互换了位置，延伸方向不同。无论哪种图，都是将不同数据集所对应的数据沿 X 轴的标签分组，以便于各组数据通过图形进行更清晰的对比。

不同的数据集可采用不同的颜色或模式表示。

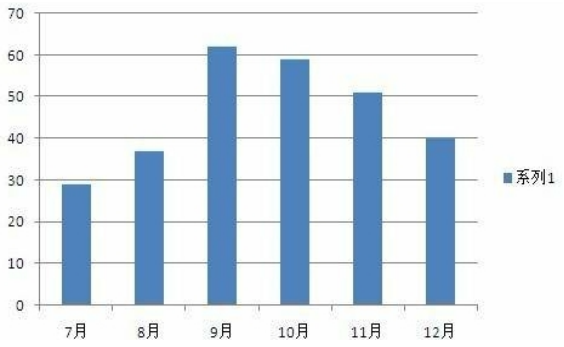


图 32 柱状图 [135]

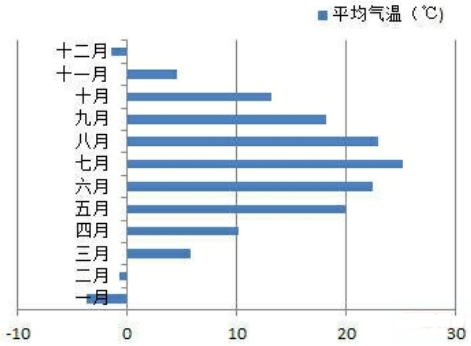


图 33 条形图 [136

(3) 散点图 (Scatter Graph)

散点图的典型用途是比较成对的数据点，其可将数据集中的每一条记录映射为二维或三维坐标系中的图形实体；在建立于坐标系的基础上，可以用点、十字叉或棍图等来表示二维坐标上的点。散点图通常被用来观察二维数据的分布情况，同时也是构建散点图矩阵的基本元素，还可以和图标显示技术相结合，很容易扩展为多维数据的可视化。另外，散点图也是最常用的数据挖掘可视化工具，可以帮助用户寻找聚类、孤立点、趋势和相互关系。

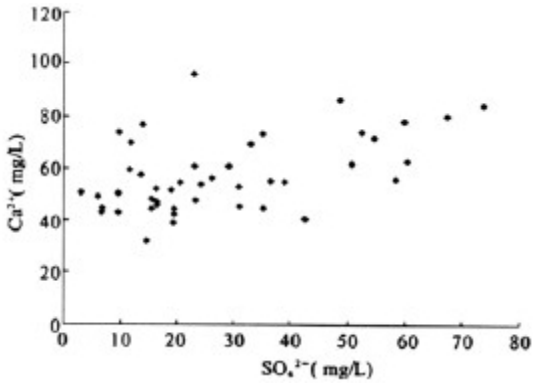


图 34 散点图 [137]

(4) 饼图 (Pie Graph)

饼图主要用于显示各种情况占总份额的分布信息。其中，离散字段的值作为饼图中每一个切片的标签，连续字段的值作为每个离散字段的值上的分组汇总，进而形成分布信息。

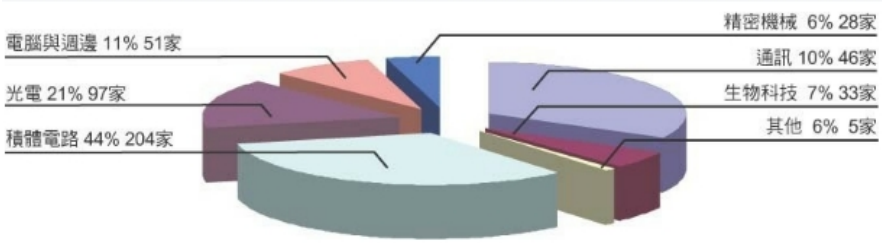


图 35 饼图 [138]

当今国际上流行的数据可视化技术主要有[139]：

11.3.2 几何显示技术

基于几何方法的多维数据可视化技术通过几何画法或几何投影的方式来表示数据库中的数据，以线或折线来表示数据中各变量之间的关联，其目标在于发现多维数据集中令人感兴趣的投影，从而将对多维数据的分析转化为仅对感兴趣的少量维度数据的分析。这种技术主要适用于数据量不大但维数较多的数据集，其优点在于比较容易观察数据的分布并发现其中的奇异点。基于几何的可视化技术主要包括散列图（Scatter Plots）、超盒图（Hyperbox）、平行坐标（Parallel Coordinates）、地形图（Landscapes）、映射追踪（Projection Pursuit）等方法。

(1) 散列图 (Scatter Plot)

散列图是一种多维数据可视化的方法。其将多维数据的各个变量两两对应，绘制该数据在二维上的分布图，从而得到一个数据的散列阵（Scatter Plot Matrix）；多维数据的每两个变量对应的分布图都作为散列阵中的一个元素（亦称为面板），进而从各个属性的两两比较中获取隐含的信息。

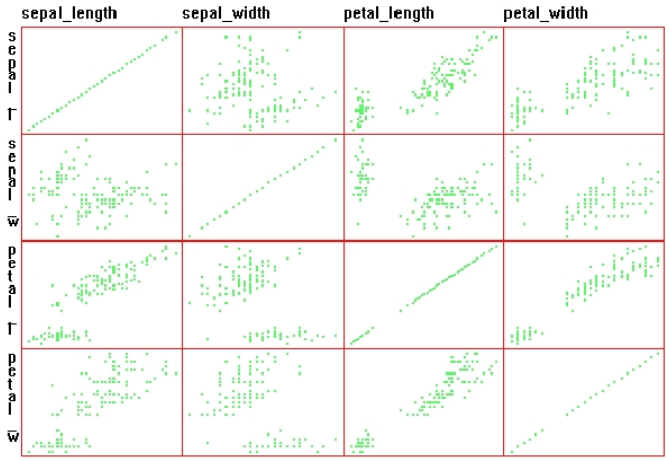


图 36 散列图 [133]

(2) 超盒图 (Hyperbox)

超盒图是一种基于散列图的扩展多维可视化方法，其将散列阵的面板置于一个超盒上，每个面板有一个方向，一个 P 维的超盒图由 P×P 条直线和 0.5×P×(P-1) 个面组成；对每一条线段，存在另外 P-1 条线段与之长度和方向相同，他们共同代表一个变量，散列图上的面板则成了由不同方向的线段组成的平面。超盒图的优点是可根据需要调整线段的长度和方向，或者只选取需要的变量而忽略不重要的变量，从而更充分地显示数据。

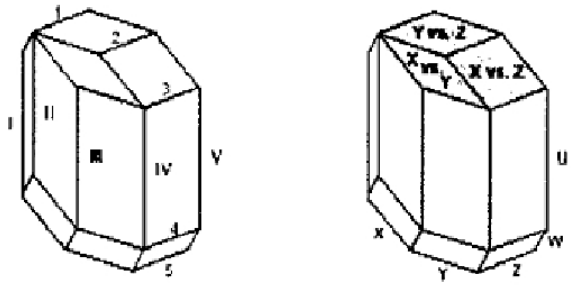


图 37 超盒图 [133]

(3) 平行坐标 (Parallel Coordinates)

平行坐标是最早提出的以二维形式表示多维数据的可视化技术之一。其基本思想是将 P 维数据空间的各属性通过 P 条等距离的平行轴映射到二维平面上，每一条轴线代表一个属性维，轴线上的取值范围从对应属性的最小值到最大值均匀分布。这样，每一个数据项都可以根据其属性值用一条折线段在 P 条平行轴上表示出来。折线顶点在坐标轴上的取值即为相应的属性取值。关系数据库的 N 个多维数据可用平行坐标上的 N 个折线来表示。这一技术能有效显示大范围的数据特性，与传统直角坐标相比，其最大优点在于：所表达的维数取决于屏幕的水平宽度而不必使用矢量或其他可视图标。但其最大的局限性在于：当数据量很

大时，大量的交迭线将使折线的密度增加，图形存在重叠，层次不清，使用户难以识别。可视化的混乱和重叠将严重阻碍用户解释可视化和彼此间的交互能力。

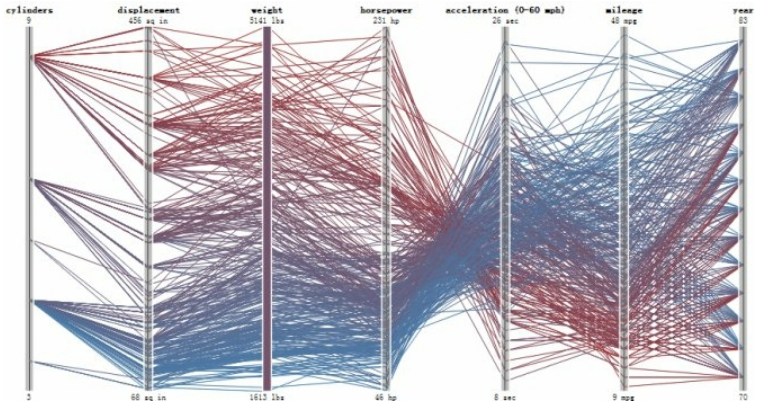


图 38 平行坐标图 [140]

11.3.3 像素显示技术

像素显示技术由德国慕尼黑大学的 D.A.Keim 提出，其基本思想是将每一个数据项的数据值对应为一个带颜色的屏幕像素，不同的数据属性采用不同的窗口分别表示，在各个单独的子窗口表现每一维的值。其优点在于可以一次性地描述大量信息并且不会产生重叠，因此，不仅能有效地保留用户感兴趣的小部分区域，也能统观全局数据。若以一个像素点表征一个数据值，则这种技术可以对目前所陈列的最大量的数据进行可视化，对于高分辨率的显示器来说，可显示多达百万数量级的数据，其主要的问题在于如何在屏幕上排列这些像素。该技术针对不同的图使用不同的排列。

(1) 递归模式技术(Recursive Pattern Technique)

递归模式技术基于一定的模式递归地生成面向行与列的排列，通过为每个递归模式所设置的参数，允许用户来控制可决定属性值排列顺序的有意义的语义结构。这一技术在分离的子窗口中可视化每一维属性，在一个子窗口中，对每一维属性值应用一个着色的像素表示，使颜色与属性值构成映射。为了使用户在同一点上将不同的属性与属性的值相关联，每一个子窗口中属性的排列顺序是相同的。该技术的特点在于按照数据属性原本存在的顺序来表现数据（如时间序列），因而适用于对多维数据集的序列分析。

(2) 圆环段技术(Circle Segments Technique)

圆环段技术针对大型高维数据集的可视化而提出，其基本思想不再是在单个子窗口表现各维属性值，而是将整个数据集以圆环的形式表现出来，每一个属性

占据圆环的一段。在这些圆环段中，属性值以单色像素表现，像素的排列从圆环的中心开始，储蓄向外围扩散。这一方法的特点在于：越靠近圆环的中心，属性就越集中，从而提高了属性值可视化的对比程度。另外，该方法不像其他面向像素的技术那么离散，因为其在中心具有一个稳固点。

11.3.4 图标显示技术

图标显示技术的基本思想是定制一些几何对象，这些三维几何对象即称为图标，然后将每一个多维数据项映射为一个图标，并按一定的顺序排列这些图标。图标的各项属性如大小、颜色、形状等均可用于与数据项的维的对应。基于图标的可视化方法包括枝形图（Stick Figures）、脸谱图（Chemoff Face）、形状编码（Shape Coding）、颜色图标（Color Icon）、表长法（Table Lens）等。图标显示技术适应于维数不多但某些维含有特别含义并且在二维平面上具有良好展开属性的数据集，用户可根据图标的显示更准确地理解这些维的含义。

(1) 枝形图 (Stick Figure)

枝形图方法的基本思想是用同一棵树枝表示多个变量，每一个变量占据一节树枝。枝形图首先选取多维数据变量中的两个变量作为基本的 X—Y 平面轴，在此平面上利用小树枝表示出其他变量值的变化，树枝的多少可以根据维数大小确定。此外，还可用树枝的颜色、粗细等特征来表示变量。

(2) 脸谱图 (Chemoff Face)

脸谱图方法是 Herman Chernoff 在 1973 年提出的一种表现多维数据的可视化方法，其希望通过脸的面部表情来呈现数据的特点。Chemoff 脸的不同部位代表不同的变量。脸谱由轮廓、眼、眼球、眉毛、鼻和嘴一共六部分组成，每一部分的长度或方向的指标可以代表变量的不同值。该方法实质上是借助人的观察能力来实现对多维数据隐含信息的探索，其将数据高度浓缩，并能反映大量数据的特征，适合在大量相似数据中发现奇异点，或者根据表情对数据进行聚类。但其缺点是图中的面部特征与数据集中数据维的对应顺序依赖关系很强，对同一组数据所产生的可视化显示不够稳定。

(3) 形状编码方法 (Shape Coding)

形状编码方法使用图标将每一维映射为一个小的像素的阵列，并将每一个数据项的像素阵列排列成正方形或矩形。对应于每一维，像素显示不同的灰度或颜

色，对应于每一数据项的正方形或矩形连续地按行排列。

(4) 彩色图标方法 (Color Icon)

彩色图标方法是一种将颜色、形状和纹理结合运用的基于图标的方法。一个彩色图标可以通过颜色、形状、大小、顺序、边界以及图标的细分来表示多维数据。绘制彩色图标一般有两种方法：一是只将区分属性的分界线描绘上映射属性值的颜色；另一是将整个细分的区域着色。

11.3.5 层次显示技术

层次显示技术的基本思想是将多维数据空间划分为若干子空间，对这些子空间仍以层次结构的方式组织并以图形表示出来。基于层次的可视化方法多利用树形结构，可直接应用于具有层次结构的数据，也可对数据变量进行层次划分，在不同层次上表示不同的变量值。该技术适用于层次关系的数据信息，例如人事组织、文件目录、人口调查等。层次显示技术主要包括树图（Tree Map）、锥形树（Cone Trees）、双曲线树（Hyperbolic Trees）、分层轴线（Hierarehiea Axis）、维堆积（Dimensional Stacking）、维嵌套（Worlds within Worlds）等方法。

(1) 树图

树图是一种屏幕填充技术，其根据属性值将屏幕进行分层分割。屏幕在 X、Y 轴方向依据属性被交替分割。树图要求须将属性值划分成类，用来分割屏幕的属性按照用户的定义排序，一般来说最重要的属性最先使用。分割出的区域的颜色可以与其它的属性相对应。

(2) 锥形树 (Cone Trees)

锥形树技术是一种三维动态的数据可视化技术，它将传统的二维树的概念扩展到了三维空间，以锥形发散形状显示树，通过颜色、形状、纹理等表现属性，通过动态缩放、手控旋转达到全面的观察效果。

(3) 双曲线树 (Hyperbolic Trees)

双曲线树是对传统树的一种变形技术，其基本思想是先将层次信息均匀地展示在双曲线面上，然后采用庞加莱（Poincare）映射方法将双曲线树映射到一个圆形区域中。因为数据对空间的要求呈指数级增长，所以双曲线技术无疑是解决这一问题的较好方法，其保留了层次的结构而只是将连接的直线弯曲成弧形，并用圆盘将整个显示界面保护起来。该技术的实现首先是在圆的中心画出树的根节

点，然后以递归的形式逐个嵌入节点。用户可以移动感兴趣的节点，将其放至圆心，双曲线技术可以通过几何变化达到平滑的动画效果。

(4) 分层轴线 (Hierarehiea Axis)

描述三维欧几里得空间的最简单方法是采用三条互相垂直的直线（例如 X 轴、Y 轴与 Z 轴）构成的笛卡尔坐标系表示，而在分层轴方法中，轴线以层次的方式水平排列。分层轴线方法一次可以在屏幕上画出的变量有限，如果数据集包含的记录数很多，超过屏幕所能显示的像素的列数，则可以采用子空间缩放技术将整个数据集的显示分布在几个面板上，并且可以将其排列成矩阵来表示。

12 总结与展望

本报告面向电能质量数据高级分析的主要问题，总结并概述了与之相关并可以应用的信息挖掘技术。

相对于面向基本目的的电能质量数据初级分析，电能质量数据高级分析的“高级”体现在以下 5 个方面：

1） 在分析目的上，由基本分析的面向统计目标转向面向预测目标；

2） 在分析目的上，从面向已经已知问题转向面向未知规律的研究；

3） 在分析范围上，从面向单个监测节点数据的分析转向面向某个区域网络的分析；

4） 在分析手段上，从单点计算为主转向面向海量数据的分布式分析技术；

5） 在分析时限上，从滞后的批量数据分析转向开始关注实时分析数据的需求。

本文针对上述挑战性问题，分别阐述了可用之于解决高级分析的相应技术，共计有 7 个方面：

1）针对电能质量监测采集的大数据，提出 分布式数据挖掘的技术框架；

2）针对电力系统产生流数据的特点，提出使用流计算框架；

3）针对电能质量问题的预测要求，归纳总结了统计回归与时间序列分析技术；

4）针对复杂决策需要，归纳了分类技术，并介绍了 9 种典型的分类算法；

5）针对发现未知分布的需要，归纳了聚类技术，并介绍了 5 类典型的聚类算法；

6）针对在局域中快速定位问题的需要，提出并联规则分析方法；

7）针对高级数据分析过程中人机交互需要，归纳总结了数据可视化技术。

在面向区域电网进行节点间影响关系分析方面，本研究提出两种创新性的方法：

1）将关联规则分析法应用于多相邻节点在某方面电能质量问题上的相互影响关系分析。其实现的挑战性问题是在现有数据基础上，生成多节点在被观察问题上的时间轴对齐的数据，以便应用现有数据挖掘算法。本文对来自实际系统的数据进行了实验分析，采用的关联规则算法为 FP-growth 算法。

2）提出一种利用全网电能质量监测数据，通过计算各监测节点之间出现谐波的顺序关系来定位谐波源的方法。在当前电能质量监测仪广泛部署应用的情况下，该方法利用监测数据记录时间连贯的特性，以迭代扫描方式从中提取各节点的谐波序列，再计算两两节点间包含 5 种预定义的序列关系的比例，进而定位出疑似谐波源位置。以生产系统中数据实例为验证，本文方法定位的谐波源与生产环境实际所知谐波源位置一致。因而，该方法在不进行额外仪器设备投入的前提下，利用历史数据，能够实现快速有效定位电网中谐波源的目的。

本文在研究过程中发现，许多应该进一步深入分析的问题，如：电压暂降原因分析、谐波源分布分析等，由于收集的监测数据没有相应的数据内容或数据关联，致使基于数据挖掘的分析方法无法进一步应用。因而，在未来的工作中，可以从数据分析的角度出发，对电能质量监测数据的采集和相关信息的收集工作提出建议，以便可以更好的进行电能质量高级数据分析。

综上所述，本文给出可能应用于电能质量数据高级分析的技术汇总与梳理，也提出了面向区域电网进行整体分析的两种新方法。此外，本研究还存在着进一步的细化和拓展创新性应用的空间。

参考文献

[1] 肖国春,刘进军,王兆安.电能质量及其控制技术的研究进展[J].电力电子技术, 2000, 34(6): 58-60

[2] 杨进,肖湘宁,王宏(Yang Jin ,Xiao Xiangning ,Wang Hong) . 网络型电能质量监测系统中 PQDIF 的实现(Realization of PQDIF in web-based power quality monitoring system) [J] . 现代电力(Modern Electric Power) ,2004 ,21 (6) :24 - 28.

[3] 杨进,肖湘宁,王宏(Yang Jin ,Xiao Xiangning ,Wang Hong) . 网络型电能质量监测系统中 PQDIF 的实现(Realization of PQDIF in web-based power quality monitoring system) [J] . 现代电力(Modern Electric Power) ,2004 ,21 (6) :24 - 28.

[4] 肖湘宁，电能质量科技发展新动态，第三届中国设计师网－电能质量高峰论坛，2009 年 8 月，北京

[5] [1]IEEE,std,1159-195.IEEE recommended practice for monitoring electric power quality.New York:IEEE,1995，1-36

[6] 林海雪.现代电能质量的基本问题.电网技术,2001,25(10):5-12

[7] 肖湘宁,徐永海.电能质量问题剖析.电网技术,2001,25(3):66-69

[8] 韩英铎,严干贵,姜齐荣等.信息电力与 FACTS 及 DFACTS 技术.电力系统自动化,2000(19):1-7

[9] Douglas J.Solving problems of power quality.EPRI Journal,1993,18(8):6-15

[10] 欧阳森.低压配电系统中电能质量监测的信号处理方法.[西安交通大学博士学位论文].西安:西安交通大学电气工程系,2003,1-2

[11] IEEE Std 1159. 3D. Recommended Practice for the Transfer of Power Quality Data [S].

[12] 聂晶晶，许晓芳，夏安邦，等. 电能监测与管理系统

[13] 许晓芳，张建平. 电能质量多数据源兼容技术简介. 电能质量国际研讨会, 2002.

[14] Christopher J. C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition". Data Mining and Knowledge Discovery 2:121 - 167, 1998

[15] KUO K， RABBAH，R. A productive programnning environment for stream computing[J]. Computer Science and Aritificial intelligence Laboratory，vol 6， no10，2005:198-201.

[16] 杨学军，曾丽芳，邓宇. Imagine 流处理器上流的优化组织方法[J]. 计算机学报，7(31)，2008: 1092-1100.

[17] S4: Distributed Stream Computing Platform [DB/OL].(2012-12-28).

[18] Puma [DB/OL].(2012-12-28). http://www.facebook.com/Puma.

[19] Twitter Storm [DB/OL].(2012-12-28). https://github.com/nathanmarz/storm.

[20] S4: Distributed Stream Computing Platform [DB/OL].(2012-12-28).

[21] Jonathan Leibiusky， Gabriel Eisbruch， Dario Simonassi. Getting Started with Storm[M]. O'Reilly Media， 2012：72-75.

[22] 郑钟光.多元回归分析在矿石体重测定中的应用[J].冶金地质动态,1986:44-46.

[23] 张保国等.山东省老年人群血压的社会影响因素多元回归分机[J].中国老年学杂志,2008, 28: 1173-1174.

[24] 刘伟铭等.基于多元回归分析的事件持续时间预测[J].公路交通科技,2005,22(11): 126-129.

[25] 徐海量等.塔电木河下游坏境因子与沙漠化关系多元回归分析[J].干旱区研究-2003,20(1): 39-43.

[26] Billing, D & J.S.Yang. Application of the ARIMA Models to Urban Roadway Travel Time Prediction - A Case Study[C]. IEEE International Conf, in SMC' 06. Taipei, 2006. pp.2529-2534.

[27] Sadek N, A.Khotanzad & T.Chen. ATM Dynamic Bandwidth Allocation Using F-ARIMA Prediction Model[C], In Proc. IEEE International Conf, in ICCCN 2003. pp.359-363.

[28]孙靖等.基于季节性时间序列模型的空调负荷预测[J].电工技术学报,2004,19(3):88-93.

[29]张熙等,含有周期性的时间序列中随机性缺失数据的填补方法[J].中国卫生统计. 2012.8 (29): 475-477.

[30]颜金木.基于趋势外推法的电力负荷预测[J].中国外资.2011:228.

[31]高志刚等.趋势外推法在地表沉降预测中的应用[J].路基工程,2010,(4):128-130.

[32]刘思峰.灰色系统理论及其应用[M].北京市:科学出版社,2010

[33]刘思峰等.GM(1,1)模型的适用范系统工程理论与实践.2000,(5):121-124.
[34]姚天样等.离散 GM(1,1)模型的特征与优化[J].系统工程理论与实践.2009,29(3):142-148.
[35]阳春华等.焦炉配煤专家系统的定性定量综合设计方法[J].自动化学报.2000,26(2): 226-232.
[36]彭怀午等.基于人工神经网络的风电场短期功率预测[J].太阳能学报.2011,32(8):1245-1250.
[37]郑鹏辉.基于 ARIMA 模型的组合模型研究[D].燕山大学.2009.
[38] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science. 1982.11 (5): 341-356.
[39] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data[M]. Dordrecht: kluwer Academic Publishers, 1991.
[40] Slowinski R. Intelligent Decision Support: Handbook of Application and Advances of Rough Sets Theory[M]. Dordrecht: kluwer Academic Publishers, 1992.
[41] Lin.T.Y & N.Cercone. Roughs and Data Mining: Analysis of Imprecise Data[M]. USA:Kluwer Academic Publisher,1996.
[42] Slezak D, G.Y.Wang et al. Rough Sets. Fuzzy Sets, Data Mining, and Granular Computing[M]. Germany: Springer-Verlag, 2005.
[43] 雷绍兰等,电力负荷的模糊粗糙集预测方法研究[J].高电压技术.2004,30(9):58-61.
[44] 张宏刚等.基于气象因素粗糙集理论的负荷预测方法[J].电力系统及其自动化学报.2004(4): 59-63.
[45] 钟波等.粗糙集与神经网络的电力负荷新型预测模型[J].系统工程理论与实践.2004(6): 113-119.
[46] 费胜巍等.融合粗糙集理论与灰色理论的电力变压器故障预测[J].中国电机工程学报.2008, 28(16): 154-160.
[47] 张诚等.基于粗糙集和多元回归分析的江西铁路物流需求预测[J].企业经济.2012(1): 112-114.
[48] 罗仁吉.单井投资估算与经济效益评价[D].西安建筑科技大学.2011.
[49] 刘盾等.基于粗糙理论的线性回归方法及实证分析[J].统计与信息论坛.2003,
[50] 夏丽. 基于ARIMA模型及回归分析的区域用电量预测方法研究.[南京理工大学硕士学位论文].南京:南京理工大学,2013.3
[51] Tso G.K.F & K.K.W.Yau. Predicting electricity energy consumption-A comparsion of regression analysis, decision tree and neural network [J], Energy, 2007,32: 1761-1768.
[52] A.Z.Al-Gami, S.M.Zubair, J.S.A.Nizami. A Regression Model for Electric Energy Consumption Forecasting in Eastern Saudi Arabia[J], Energy, 1994: 1043-1049.
[53] 谢龙汉,尚涛.SAS 统计分析。数据挖掘[M].北京:电子工业出版社.2012:230-259.
[54] 王振龙.时间序列分析[M].北京:中国统计出版社,2000:1-27.
[55] 徐兴梅等.时间序列分析关键问题研究[J].农业网络信息.2010(1): 48-50.
[56] Box G.E.P, G.M.Jenkins, GC.Reinsel.时间序列分析、预测与控制[M].王成璋等,译.北京:机械工业出版社,2011: 1-68.
[57] 刘瑛慧等.时间序列分析理论与发展趋势[J].电脑知识与技术.2010,6(2): 257-258.
[58] Box G.E.P, G.M.Jenkins, GC.Reinsel.时间序列分析一预测与控制[M].王成璋等,译.北京:机械工业出版社,2011: 1-68.
[59] 刘瑛慧等.时间序列分析理论与发展趋势[J].电脑知识与技术.2010, 6(2): 257-258.
[60] 林宇,等. 数据仓库原理与实践[M]. 北京:人民邮电出版社, 2005.
[61] Guha S, Rastogi R, Shim K. CURE: An Efficient Clustering Algorithm forLarge Databases[C]. Seattle: Proceedings of the ACM SIGMOD Conference, 1998. 73284.
[62] Guha S, Rastogi R, Shim K. ROCK: A Robust Clustering Algorithm for Categorical Attributes[C]. Sydney: Proceedings of the 15 th ICDE, 1999. 5122521.
[63] Karyp is G, Han E2H, Kumar V. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling [J]. IEEE Computer,1999, 32 (8) : 68-75.
[64] Ester M, Kriegel H2P, Sander J, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [C].Portland: Proceedings of the 2nd ACM SIGKDD, 1996. 226-231.

[65] Hinneburg A, Keim D. An Efficient Approach to Clustering Large Multimedia Databases with Noise[C]. New York: Proceedings of the 4th ACM SIGKDD, 1998. 58-65.
[66] Wang W, Yang J, Muntz R. STING: A Statistical Information Grid Approach to Spatial Data Mining [C]. Athens: Proceedings of the 23rd Conference on VLDB, 1997. 186-195.
[67] Wang W, Yang J, Muntz R R. STING+ : An Approach to Active Spatial Data Mining [C]. Sydney: Proceedings of the 15th ICDE,1999. 116-125.
[68] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications[C]. Seattle: Proceedings of the ACM SIGMOD Conference, 1998.94-105.
[69] Sheikholeslami G, Chatterjee S, Zhang A. WaveCluster: A Multiresolution Clustering App roach for Very Large Spatial Databases [C].New York: Proceedings of the 24 th Conference on VLDB, 1998. 428-439.
[70] Chris Ding. A Tutorial on Spectral Clustering[C]. ICML, 2004.
[71] Mitchell T. Machine Learning[M]. New York: McGraw2Hill, 1997.
[72] Ertoz L, SteinbachM, KumarV. Finding Clusters ofDifferent Sizes, Shapes, and Densities in Noisy, High Dimensional Data [R].Minneapolis: University of Minnesota, 2002.
[73] Kaufman L, Rousseeuw P. Finding Group s in Data: An Introduction to ClusterAnalysis[M]. New York: JohnWiley and Sons, 1990.
[74] Ng R, Han J. Efficient and Effective ClusteringMethods for Spatial DataMining[C]. Santiago: Proceedings of the 20 th Conference on VLDB, 1994. 144-155.
[75] Bradley P, Fayyad U. Refining Initial Points for K2means Clustering[C]. Madison: Proceedings of the 15 th ICML, 1998. 91-99.
[76] Dhillon I, Guan Y, Kogan J. Refining Clusters in High Dimensional Data[C]. Arlington: The 2nd SIAM ICDM, Workshop on Clustering High Dimensional Data, 2002.
[77] Zhang B. Generalized K-harmonic Means: Dynamic Weighting of Data in Unsupervised Learning[C]. Chicago: Proceedings of the 1st SIAM ICDM, 2001.
[78] PellegD,Moore A. X-means: Extending K-means with Efficient Estimation of the Number of the Clusters[C]. Proceedings of the 17 th ICML, 2000.
[79] Sarafis I, Zalzala A M S, Trinder PW. A Genetic Rule-based Data Clustering Toolkit[C]. Honolulu: Congress on Evolutionary Computation (CEC) , 2002.
[80] Strehl A, Ghosh J. A Scalable Approach to Balanced, High-dimensional Clustering of Market Baskets[C]. Proceedings of the 17th International Conference on High Performance Computing, Bangalore:Springer LNCS, 2000. 525-536.
[81] Banerjee A, Ghosh J. On Scaling up Balanced Clustering Algorithms[C]. Arlington: Proceedings of the 2nd SIAM ICDM, 2002.
[82] Banerjee A, Ghosh J. On Scaling up Balanced Clustering Algorithms[C]. Arlington: Proceedings of the 2nd SIAM ICDM, 2002.
[83] Tung A K H, Hou J, Han J. Spatial Clustering in the Presence of Obstacles[C]. Heidelberg: Proceedings of the 17th ICDE, 2001. 359-367.
[84] Han J, KamberM, TungA K H. Spatial Clustering Methods in Data Mining: A Survey[C]. Geographic Data Mining and Knowledge Discovery, 2001.
[85] Kohonen T. Self-Organizing Maps [M]. Springer Series in Information Sciences, 2001. 30.
[86] Yongqiang Cao, Jianhong Wu. Dynamics of Projective Adaptive Resonance Theory Model: The Foundation of PART Algorithm [J]. IEEE Transactions on Neural Network, 2004, 15 (2) : 245-260.
[87] Brown D, Huntley C. A Practical Application of Simulated Annealing to Clustering[R]. University of Virginia, 1991.
[88] Cristofor D, Simovici D A. An Information Theoretical Approach to Clustering Categorical Databases Using Genetic Algorithms [C]. Arlington: The 2nd SIAM ICDM, Workshop on Clustering High Dimensional Data, 2002.
[89] Ganti V, Gehrke J, Ramakrishna R. CACTUS-Clustering Categorical Data Using Summaries[C]. San Diego: Proceedings of the 5th ACMSIGKDD, 1999. 73-83.
[90] Dhillon I. Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning [C]. San Franciso: Proceedings of the 7th ACM SIGKDD, 2001. 269-274.

[91]林海雪.现代电能质量的基本问题.电网技术, 2001, 25(10):5-12
[92]韩英铎,严干贵,姜齐荣等.信息电力与 FACTS 及 DFACTS 技术.电力系统自动化,2000(19):1-7
[93]肖湘宁,徐永海.电能质量问题剖析.电网技术,2001,25(3):66-69
[94]欧阳森.低压配电系统中电能质量监测的信号处理方法.[西安交通大学博士学位论文].西安:西安交通大学电气工程系,2003,1-2
[95] Han J W,Kamber M.Data mining: concepts and techniques[M].Morgan Kaufmann,2005.
[96] Han jia wei, Pei Jan 等 Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach.2004
[97] 肖湘宁主编.电能质量分析与控制[M].北京:中国电力出版社,2004 年
98 Xu W.Power Direction method cannot be used for harmonic source detection. Power engineering society summer meeting.2000.IEEE, Volume 2, 16-20 July 2000
99 张庆河.电网谐波源的检测与定位[J].山东电力技术,2005,5:71-73
[100] Chun Li,Wilsun Xu,Tayjasanant.T.A“critical impedance” based method for identifying harmonic sources[J].IEEE Transactions on Power Delivery,Volume.19,Issue.2,April 2004
[101] Ahmed A M ,Abdel M ,Mahmud A E .Separation of Customer and Supply Harmonics in Electrical Power Distribution System Ninth International Conference on Quality of Power Proceedings .Orlando(FL):2000
[102] Liu Yamei,Gong Hualin,Xiao Xianyong,Yang Honggeng.Harmonic Source Location at the Point of Common Coupling Based on the Nonlinearity Index of Load[C].Power and Energy Engineering Conference,March,2009,1-5
[103] 田立亭,程林,孙元章等.用户侧谐波源对 PCC 谐波水平的影响与区分[J].继电器,2007,21：59-63
[104] Kazimierz Wilkosz.Harmonic Sources Localization: Comparison of Methods Utilizing the Voltage Rate or the Current Rate [J].IEEE EPQU, 2007, 1-8
[105] Kazimierz Wilkosz.A Generalized Approach to Localization of Sources of Harmonics in a Power System [J].IEEE ICHQP, 2008, 1-6
[106] Kazimierz Wilkosz.Localization of Harmonic Sources in a Power System with Use of the Generalized Localization Rate [J].IEEE EPQU, 2009, 9, 15-17
[107] 吴竞昌(Wu Jingchang) . 供电系统谐波(Harm onics in Power Supply System) [M] . 北京:中国电力出版社(Beijing : China Electric Power Press) , 1998.
[108] Task Force on Harmonics Modeling and Simulation. Modeling and simulation of the propagation of harmonics in electric power networks, Part I: Concepts, models and simulation techniques[J] . IEEE Trans on Power Delivery , 1996, 11(1) : 452- 465.
[109] Task Force on Harmonics Modeling and Simulation.Characteristics and modeling of harmonic sources power electronic devices [J]. IEEE Trans on Power Delivery, 2001, 16(4) : 791- 800.
[110] Hiyama T, Hamman M S A A, Ortmeyer T H. Distribution system modeling with distributed harmonic sources [J] . IEEE Trans on Power Delivery, 1989, 4(2) : 1297- 1304.
[111] Thunberg E, Soder L. A norton approach to distribution network modeling for harmonic studies [J] .IEEE Trans on Power Delivery , 1999, 14(1) : 272-277.
[112] Fauri M. Harmonic modeling of non-linear load by means of crossed frequency admittance matrix [J] .IEEE Trans on Power System , 1997, 12(4) : 1632-1638.
[113] 赵勇, 张涛, 李建华, 等(Zhang Yong , Zhang Tao , Li Jianhua, et al) . 一种新的谐波源简化模型(A new simplified harmonic source model for harmonic analysis and mitigation) [J] . 中国电机工程学报(Proceedings of the CSEE) , 2002, 22(4) : 46-51.
[114] 殷桂梁, 容亚君, 肖丽萍(Yin Guiliang , Rong Yajun, Xiao Liping) . 非线性负载的谐波模型研究(The research of harmonic model of the nonlinear load) [J] .电力系统及其自动化学报(Proceedings of the CSUEPSA), 1998, 10(3) : 58-61.
[115] 吴笃贵, 徐政(Wu Dugui, Xu Zheng) . 电力负荷的谐波建模(Harmonic modeling of electric load)[J] . 电网技术(Power System Technology), 2004, 28(3) : 20-24.
[116] 王茂海, 刘会金(Wang Maohai, Liu Huijin) . 通用瞬时功率定义及广义谐波理论(A universal definition of instantaneous power and broad-sense harmonic theory) [J] . 中国电机工程学报(Proceedings of CSEE) , 2001, 21(9) : 68-73.

[117] Heydt G T. Identification of harmonic source by a state estimation technique[J] . IEEE Trans on Power Delivery, 1989, 4(1) : 569- 576.
[118] Hong Y Y, Chen Y C. Application of algorithms and artificial-intelligence approach for locating multiple harmonics in distribution systems[J].IEEE Proceedings Generation, Transmission and Distribution,1999, 146(3) : 325-329.
[119] Teshome A. Harmonic source and type identification in a radial distribution system [A]. In: Proceedings of IEEE Industry Applications Society Annual Meeting[C].Michigan,USA: 1991. 1605-1609.
[120] Tanaka T , Akagi H. A new method of harmonic power detection based on the instantaneous active power in three-phase circuits [J] . IEEE Trans on Power Delivery , 1995, 10(4) : 1737-1742.
[121] Aiello M , Cataliotti A, Cosentino V, et al . A selfsynchronizing instrument for harmonic sources detection in power systems[A] . In: Proceedings of the 20th IEEE Instrumentation and Measurement Technology Conference [C] , Colorado , USA: 2003. 1364-1369.
[122] Dan A M, Czir a Z. Identification of harmonic sources[A] . In: Proceedings of the 8th International Conference on Harmonics and Quality of Power[C] , Athens,Greece : 1998. 831-836.
[123] Moustafa A A, Mo ussa A M , El-Gam mal M A. Separation of customer and supply harmonics in electrical power distribution systems [A] . In: Proceedings of the 9th International Conferenceon Harmonics and Quality of Power [C] , Orlando, USA: 2000.1035-1040.
[124] 李晓梅. 科学计算可视化导论. 北京：国防科技大学出版社, 1998, 112-209.
[125] David McCandless. 数据可视化之美.
[126] 任永功. 面向聚类的数据可视化方法及相关技术研究. 沈阳:东北大学, 2006.
[127] 郑昌璇, 陈洋. 大数据下可视化分析. 技术研发. 2013, 20(6): 32-34.
[128] 雷琴琴. 面向像素的可视化技术研究. 北京: 北京交通大学, 2007.
[129] Keim D A. Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics, 2002, 7(1): 1-8.
[130] 刘俊霞. 数据聚类及可视化技术. 成都: 电子科技大学, 2008.
[131] L. Nowell, S. Havre, B. Hetzler, P. Whitney. Themeriver: Visualizing. Thematic Changes in Large Document Collections, IEEE Transactions on Visualization and Computer Graphics, 2002, 8(1): 9-20.
[132] N. Lopez, M. Kreuseler, H. Schumann. A Scalable Framework for Information Visualization, IEEE Transactions on Visualization and Computer Graphics, 2002, 8(1): 39-51.
[133] 于洋. 数据挖掘可视化技术的研究与应用. 长春：吉林大学, 2008.
[134] 常州楼市 1-10 月销售排名情况报告. [DB/OL]. http://cz.house.sina.com.cn/scan/2010-11-09/ 154514355.shtml.
[135] 2011 年底最后一搏重庆楼市 12 月 40 余盘开盘. [DB/OL]. http://cq.fang.anjuke.com/news/ 2011-11-22/137475.html.
[136] 在 Excel2010 图中用指定颜色表示负值. [DB/OL]. http://office.wps.cn/officeexcel/ 29553-2013-04-10-17-56-48-54.html.
[137] 鄂尔多斯市哈头才当水源地水质评价及保护措施. [DB/OL]. http://www.hwcc.gov.cn/pub/ hwcc/wwgj/ bgqy/mssp/201110/t20111018_337064.html.
[138] 产业别圆饼图图片. [DB/OL]. http://www.nipic.com/show/3/73/3599852ke837b099.html.
[139] 胡俊. 数据挖掘可视化模型及其应用研究. 北京: 北京交通大学, 2009.
[140] 图可视化工具“Protovis”介绍及经典案例. [DB/OL]. http://blog.sina.com.cn/s/blog_ 6355dfc60101344g.html.