北京林业大学

学术型硕士生学位论文开题报告

题　　目： 基于频域变换的森林生态站观测指标聚类分析

学　　号	3130295	**姓　　名**	张兆玉
学科专业	计算机软件与理论	**研究方向**	数据挖掘与机器学习
导师姓名	陈志泊	**职　　称**	教授
	王建新		教授
报告主持人		**报告日期**	2014.09.16

填表日期： 2014 年 09 月 10 日

填 表 说 明

1、开题报告是硕士生培养的重要环节，研究生需在导师的指导下认真完成，具体要求参见《北京林业大学关于学术型研究生开题报告的规定（修订）》。

2、开题报告文献综述部分的基本要求：（1）国内外本研究课题的发展现状、趋势及问题等，字数 6000 字左右。（2）参考文献量不少于 30 篇（其中人文社科类不少于 40 篇），对于个别新兴研究领域其文献量可酌情减少。（3）文献引用格式需符合《北京林业大学研究生学位论文格式的统一要求》的相关规定。

3、完成时间：研究生开题工作应于入学后第三学期内完成，具体时间各学院可根据本学院的学科特点和实际情况进行安排。

4、硕士生开题报告书应首先获导师认可和考核小组成员审阅后方可参加开题。

5、打印要求：此表用 A4 纸双面打印，各栏空格不够时，请自行加页。

6、开题报告通过、修改、签字完毕后，交各学院存档一份。

选题基本情况（ √ ）

本研究题目为：

1. 导师课题的一部分 (√)；

2. 委培单位的课题 ( )；

3. 其它（须具体说明）_____。

选题分类 （ √ ）

1. 基础研究 (√)　　2. 应用研究 ( )

3. 综合研究 ( )　　4. 其　他 ( )

选题来源 （ √ ）

1. 973、863 项目( )　　2. 国家社科规划、基金项目 ( )

3. 教育部人文、社会科学研究项目( )　　4. 国家自然科学基金项目 ( )

5. 中央、国家各部门项目 ( )　　6. 省（自治区、直辖市）项目( )

7. 国际合作研究项目( )　　8. 与港、澳、台合作研究项目（ ）

9. 企、事业单位委托项目（ ）　　10. 外资项目（ ）

11. 学校自选项目 （ √ ）　　12. 国防项目 （ ）

13. 非立项 （ ）　　14. 其他 （ ）

一、立题依据

1. 选题的理论和实践意义

为了在更大的范围和尺度上揭示生态系统的变化规律，减少生态系统管理的不确定性，长期的生态系统联网研究和监测是一种有效的方法。八十年代以来世界上建立了多个生态系统研究网络。在国家级别上，美国的长期生态研究网络（LTER）、英国的环境变化研究监测网络（ECN）、加拿大的生态监测与分析网络（EMAN）、中国生态系统研究网络(CERN)[1-2]。

中国的生态系统研究网络（CERN）于 1988 年开始组建成立。目的是为了监测中国生态环境变化，综合研究中国资源和生态环境方面的重大问题，发展资源科学、环境科学和生态学。目前，该研究网络由 13 个农田生态系统试验站、9 个森林生态系统试验站、2 个草地生态系统试验站、6 个沙漠生态系统试验站、1 个沼泽生态系统试验站、2 个湖泊生态系统试验站、3 个海洋生态系统试验站，以及水分、土壤、大气、生物、水域生态系统 5 个学科分中心和 1 个综合研究中心所组成。

森林作为人类文化的摇篮和绿色宝库，是重要的可再生资源，无疑使得森林生态站的存在举足轻重。我国森林生态系统试验站成立后，确定了长期的研究方向：(1) 森林生态系统的结构与功能；(2) 山地生态系统多样性的保护和持续利用的途径；(3) 退化生态系统自然演替规律以及恢复途径和人工优化生态系统的组建；(4) 全球变化对各类生态系统结构、功能和动态过程的影响。通过几十年对研究方向的专注和探索，我国的森林生态站的发展已经取得长足的发展，并积累了大量丰富的观测数据。发展至今，我们面临的问题是该如何利用这些宝贵的数据，来发现其潜在的价值[3]。

数据挖掘[4-6]（Knowledge Discovery in Data 简称：KDD）技术的出现为我们从浩如烟海的数据中挖掘黄金提供了可能。利用数据挖掘的一般过程：数据清理、数据集成、数据选择、数据变换、数据挖掘、模式评估，到最终发现有用的知识（即模式），我们可以对生态站的数据进行科学有效的分析。数据挖掘的功能主要包括：(1) 类/概念描述：特征化与区分；(2) 挖掘频繁模式、关联与相关性；(3) 用于预测分析的分类与回归；(4) 聚类分析；(5) 离群点监测。

通过对现有的森林生态站观测指标数据的预处理和分析，发现如下问题：

(1) 缺乏一个统一的平台来对观测的数据进行管理，导致数据的价值没有被最大化利用。数据的共享严重不足。

(2) 由于观测仪器老化、断电、系统故障等原因导致数据缺失，是否可以根据数据之间的相关性，进行数据补全工作。

(3) 观测指标的观测值是基于时序的，各观测指标间是否存在某种时间周期上的相关性。

针对上述问题，我们确立了课题研究方向：构建统一的森林生态站观测数据管理平台，旨在实现数据的自动采集、传输、存储、数据分析和处理，达到数据的标准化、规范化的管理，为研究人

员进行研究提供有力的平台支持；象限近邻填充算法（Quadrant Encapsidated Nearest Neighbor Based Imputation 简称：QENNI）于多维空间上的数据补全[13-21]与基于离散时间傅里叶变换[22-24]（Discrete-time Fourier Transform 简称：DTFT）的森林生态站观测指标的聚类分析；由于生态站观测指标数据是在时域上变化的，一般很难看出其是否具有周期规律。我们尝试采用 DTFT 将时域上的函数转变为频域上的函数，然后利用距离公式计算观测指标间的距离，拟通过该方式查看各个指标间的内在联系。k 最近邻填充算法（kNN）中选取的 k 个最近邻点有偏好，而 QENNI 仅仅使用缺失数据象限方向的最近邻数据填充该缺失值，其对于低维数据集可以是无参的，即消除了对参数的依赖。

通过将基于离散时间傅里叶变换和 QENNI 算法引入到森林生态指标数据的分析中，不仅可以揭示各个指标间内在的联系，为下一步数据的分析奠定基础。同时，通过 QENNI 数据补全算法能够保证观测指标数据的正确性和准确性，也为聚类算法建立模型提供了精准的数据基础。

2. 文献综述（国内外本研究领域的发展现状、趋势及问题等，并附参考文献）

随着国内外森林生态观测站的建立和完善，其信息化管理需求更加突出。互联网的高速发展，“信息化”和“数字林业”的概念早已提出。森林生态站走向信息化道路是必然的选择。观测站几十年观测的数据无疑是长久以来森林生态站建设保留下来最珍贵的资产。如果没有成熟的数据采集、管理和分析手段的有力支持，这些数据就不能被真正意义上挖掘。但是“数据丰富，信息匮乏”的现状，正是我们所面临的现实问题。随着我国对林业建设的大量投入，一些基础的森林生态站数据挖掘系统也得到研究与运用。但是，总体来说从事森林生态站数据挖掘工作的人员较少，数据中所隐藏的巨大价值还未被充分发现。可以说现在仍处在森林生态数据挖掘的起步阶段。以下分别对当前的基于大数据的数据挖掘和森林生态站数据挖掘的研究现状、趋势以及存在的问题进行阐述：

1. 数据挖掘技术研究现状

计算机技术和互联网的高速发展，直接带来了整个世界各行业的信息化革命。计算机的普及和信息化显著地改善了我们产生和收集数据的能力。特别是在数据库系统的出现之后，大量的数据从我们的生活中涌现。数据的爆炸性增长激起了对新工具和自动工具的需求，用来帮助快速地将海量数据转化为有用的信息和知识。这就导致了数据挖掘学科的产生（又称从数据中发现知识，简称 KDD）。在第 11 届国际联合人工智能学术会议上 KDD 被首次提出。截止到现在，数据挖掘领域已基本成熟。因此，KDD 国际会议研讨会的研究重点已经从方法过度到应用。

在国外，数据挖掘的典型应用早已出现。例如在商务智能（BI）领域，很好地把握诸如顾客、市场、供应和资源以及竞争对手的状况等是十分重要的。BI 能够提供商务运作的历史、现状和预测

视图，包括业绩管理、标杆管理和预测分析。特别是在电子商务领域的用户个性化推荐中，也是数据挖掘的典型应用场景；商场从顾客购买商品中发现一定的关联规则，提供打折、购物券等促销手段，提高销售额；在制造业中，半导体的生产和测试中都产生大量的数据，就必须对这些数据进行分析，找出存在的问题，提高质量；在生物领域，采用数据挖掘的手段对 DNA 进行分析；在银行和保险行业中通过离群点检测对发生的异常行为进行检测；美国 AutoTrader.com 是世界上最大的汽车销售站点，每天都会有大量的用户点击网站，寻求自己想要的信息，其运用了 SAS 软件进行数据挖掘，每天对数据进行分析，找出用户的访问模式，对产品的喜欢程度进行判断，并设特定服务，取得了成功。Reuteres 是世界著名的金融信息服务公司，其利用的数据大都是外部的数据，这样数据的质量就是公司生存的关键所在，必须从数据中检测出错误的成分。Reuteres 用 SPSS 的数据挖掘工具 SPSS/Clementine，建立数据挖掘模型，极大地提高了错误的检测，保证了信息的正确和权威性。另外，机器学习、数据库、数据仓库、人工智能、信息检索等领域的国际学术期刊业都先后开辟了数据挖掘专题或专刊。当前，国外的数据挖掘主要集中在对知识发现方法的研究，通过将不同学科的新方法的融合来显著增强数据挖掘的能力。例如，为了挖掘自然语言文本数据，将数据挖掘方法与自然语言处理和信息检索结合在一起。而在应用方面通过商业数据挖掘软件的不断产生和完善，针对问题的领域建立一个系统的整体方面是现在发展的方向。

在中国，数据挖掘的起步要晚于国外。我们可以发现数据挖掘相关的高质量的书籍更是屈指可数。但是，学习数据挖掘基础理论和应用研究的人越来越多，也使得中国的数据挖掘领域正在飞速地前进。目前，在我国清华大学、中科院计算机研究所、人民大学等在内的很多单位都已经开设数据挖掘相关的学习课程。在国内数据挖掘领域主要研究为数据挖掘方法，包括关联规则中的频繁项集、聚类、分类等；数据挖掘的应用，包括天猫、京东等电商平台的对用户数据的挖掘，百度、搜狗等搜索公司也对用户数据和日志数据进行挖掘，从中发现很多有价值的模式。在未来，数据挖掘肯定会渗透到越来越多的领域，跟行业特点紧密相关的挖掘方法肯定会越来越多地被运用。同时，中国乃至世界都会开始思考数据挖掘与社会的关系，包括数据的不适当披露和使用，个人隐私的泄露等社会问题。

2. 大数据时代的数据挖掘面临的挑战

人类自 2010 年便进入到大数据时代，在数据中生活的我们同时也在制造着数据。通过对数据的分析加快了经济的发展速度，提高了社会的文明程度。大数据不但包括大数据技术[7]、应用还包括大数据科学以及大数据工程。在大数据时代，如何深层次开发大数据并提供相关服务能力将成为竞争的关键。

大数据（Dig Data）时代的来临，给数据挖掘带来了新的机遇和挑战。大数据指不用随机分析

法（抽样调查）这样的捷径，而是采用所有的数据进行分析的方法。大数据显著的特点是种类多、流动速度快以及海量的数据。挖掘大数据库的流数据将成为一个重要的问题。这种类型的数据广泛地存在于互联网、无线通信网络、地质测量、天气、天文观测等方面，由于数据流迅速、大量、连续地到达，因此现有的数据挖掘算法在处理如此大量的数据方面速度太慢了，需要研究新的算法。与此同时，数据流需要以近实时的方式对更新流进行复杂分析[8]，这对研究者来说也是一个挑战。时间序列数据挖掘[]。时间序列是数据存在的特殊形式，序列的过去值会影响到将来值。这种影响的大小以及影响的方式可由时间序列中的趋势周期及非平稳等行为来刻画。一般来讲，时间序列数据都具有噪声、不稳定、随机性等特点，这就使得正确进行短期和长期的预测都非常困难，如何解决时间序列数据的噪声问题。从而有效地聚类、分类和预测数据趋势仍然是个有待解决的问题。过程数据挖掘一个重要的问题是如何使数据挖掘过程自动化。在数据挖掘系统里面建立一种方法来帮助用户避免许多数据挖掘中的错误。如果我们能够将各种数据挖掘过程自动化，就可以大大地减少劳力。利用目前的技术虽然可以快速地建模和寻找模式，但 90%的成本浪费在预处理上，减少这些成本将极大地降低建模的成本。另一个重要的问题是如何将可视化和自动化数据挖掘技术结合在一起，在很多应用上，数据挖掘的目标和任务不太明确，特别是在实验性数据分析上。动态数据、RFID 数据和传感器网络数据挖掘。随着传感器网络、GPS、手机和其他移动设备和 RFID 技术的普遍。大量动态数据需要被分析。在动态数据、RFID 数据和传感器数据挖掘领域里，还有许多尚未被研究的问题：例如寻找关联和规则性来清理有噪音的传感器网络数据、如何为这些数据构建数据仓库、如何对千兆字节的 RFID 数据进行挖掘、如何降噪多维轨道数据等等。大数据时代对处理数据的数量和速度都提出了新的要求。大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。因此，传统的数据挖掘方法在这些方面不能适用，需要为大数据的分析做出新的调整。分布式架构下的依托云计算、分布式数据库和云存储以及虚拟化技术的数据挖掘方法便在这样的一个背景下诞生。这是对已有的数据挖掘算法提出的新的时代要求。

在未来，大数据挖掘将成为信息安全发展的契机。随处可见的数据降低了自身信息的安全性。大量数据将会存储在云端，导致无法集中管理，这就将会引发非法入侵或者窃取、篡改数据信息的危险性提高。基于此，大数据领域研发的各种为信息安全服务的技术和产品，将保证各大数据产业链的数据安全。大数据挖掘也将成为企业及教育机构转折点。随着大数据挖掘技术在企业管理中的应用以及其带来的经济效益，企业若想在新时代的浪潮中继续前进，就必须重新定制管理模式，将大数据挖掘运用到企业生产中来。同时，大数据也将成为创造价值的核心因素。大数据中潜在的价

值是巨大的，未来的企业和政府必将都是以数据为中心的。

我们有理由相信，未来数据挖掘必定会结合云计算、大数据取得新的成功。

3. 森林生态站数据挖掘系统的研究现状

鉴于我国的数据挖掘技术起步较晚，总体的发展时间还不长。目前来说数据挖掘虽然在很多领域内得到了一定的应用，但是在国内外的文献中，森林生态站观测指标数据的挖掘技术和相关文献还十分匮乏。总体来说，数据挖掘技术在林业领域的研究也取得了相应的发展[9-10]，取得了一定的成果，但是仍然初级发展阶段。

数据挖掘技术在国内林业领域中的应用主要有：（1）基于 Apriori 算法的关联规则挖掘方法在森林资源统计中的应用。它是通过 Apriori 算法对林木的权属、起源、年龄、平均胸径、树高和郁闭度等属性之间的关联进行分析，通过该分析方法有助于评估者认识和理解其中存在着的有效知识和客观规律。（2）基于可视化技术的数据挖掘方法在营林生产统计中的应用。它是运用可视化数据挖掘技术对采种、育苗、造林等整个过程的多维信息进行综合分析，并从中发现多维信息间的复杂关系和综合影响，以模拟预测营林生产活动的整个过程和生态环境态势。（3）在林业系统中应用分类模式，大部分是对林相图等空间数据的分类，例如产生了自动绘制森林类型图的专家系统、利用人工神经网络技术构建林地因子与地位指数关系模型评价评价林地立地质量、利用 ANN 技术和 LTM（Landsat Thematic Mapper）数据进行土地分类研究等；（4）关联模式的使用实现了土地、阔叶林、针叶林等的分类，用决策树 C4.5 算法进行森林资源二类调查数据分析等；（5）使用回归模式建立林木声场模型，模拟树木残存率、评价树木生长模型等。在国外，（1）俄罗斯的 Poly Analyst 基于神经元网络技术在森林病虫害防治统计中的应用。利用人工神经网络方法计算和预测该地方害虫的发生面积和防治率。（2）美国的 Pattern Recognition Workbench 基于事例的推理方法在森林病害诊断统计中的应用。它是通过输入林木病害特征属性，确定病害特征属性特征向量和病害特征向量权值，系统寻找与现有情况相类似的事例，并选择最佳的相同的解决方案。（3）美国的 Gene Hunter 基于遗传算法在林业生产设备利用情况统计中的应用。它可以对林业生产设备的结构优化设计、设备故障诊断、温室花卉生产、智能控制等方面进行分析。这些科研成果表明在林业数据中数据挖掘技术得到了应用。

同时我们也可以看出，在国内外森林生态站中数据挖掘技术的应用往往比较单一。分类模式、关联规则等挖掘方法仅仅单一的使用在生态站数据中。综合性的挖掘方法使用还未得到系统的使用。多方法，多学科的交叉挖掘是数据挖掘新的发展方向。在未来的研究中多种方法相结合系统地运用到森林生态站数据挖掘中，可以发现很多原本隐藏的模式，用来揭示数据内在的规律，必定可以用来解决很多实际中遇到的问题。不仅可以用来分析生态站的指标周期、挖掘各观测指标之间的

内在联系，又可以聚类分析观测指标数据特点，通过得到的关联关系预测未来的变化规律，补全缺失数据等。因此，可以发现将新的挖掘手段应用到森林生态站的观测数据中进行模式发现、聚类分析将给这一领域注入新的力量。

4. 森林生态站数据管理面临的问题

随着森林生态站的发展，森林生态站的管理工作变得更加困难。主要表现在：

（1）森林生态站观测的数据比较分散。通过调查发现生态站的数据通常由不同的研究人员按自己的需要去相应的设备中采集。这种分散的采集方式造成数据的严重分散。有特殊要求的研究人员要获取自己研究所需要的数据往往需要联系很多人。这将花费很大的精力，造成了严重的人力浪费。

（2）数据量巨大，传统的文件式存储已难以满足现实的需求[11-12]。生态站的观测数据一般都是几年几十年如一日的采集而来，这将是海量的数据。传统的文件存储给数据的存储、管理、访问带来了严重的不便。管理人员要将成千上万的文件分类别、分时间管理，这需要巨大的人力投入。即使是这样也难以保证数据的大规模存储和获取。

（3）依托计算机的信息化时代早已来临，经过长期观测和研究积累下来的海量的、格式多样的观测数据的管理手段还是十分落后的，很多生态站基本上还是半手工甚至是纯手工的状态来整理管理这些数据。

（4）在当前的工作状态环境下，全国各地的生态站之间并没有建立起统一的数据管理与共享平台。各森林生态站仍就是一个个信息孤岛。这种信息交流的不通畅，阻碍了数据挖掘的数据源的多维度实现。如果能够实现一个统一的平台，我们相信运用数据挖掘将会发现更多有价值的模式。

（5）随着数据量的增加，新的应用情景下需要新的数据挖掘算法。传统的单一算法单一分析已经不能满足需求。多维度，多次迭代的挖掘方法是新的情景下的新要求。

5. 森林生态站数据挖掘的趋势

数据挖掘技术正在以前所未有的速度发展着，已经被运用到我们生活的很多方面。通过整体的数据挖掘领域的发展，数据挖掘技术在森林生态站领域[25-26]的应用将出现以下趋势：

（1）统一的多站合作的数据共享管理平台的建立。面对当前相对分散的数据管理现状，建立统一的多站数据共享管理平台能够增加可分析数据的维度和广度。这将在很大程度上促进新的模式和规律的发现。同时，也将给从事森林生态站管理工作和研究的人员带来极大的便利。

（2）基于分布式存储的大数据存储方案[27-32]运用到森林生态站数据管理平台。现有的落后的数据管理方式已经给森林生态站的发展带来了瓶颈。通过建立分布式的数据存储方案，将给依赖于数据的云计算和大数据挖掘提供基础。

（3）多种算法相结合的数据挖掘方法[33]将被更多的应用到森林生态站数据的挖掘工作中。通过将其他行业成功的算法引入到生态站的数据挖掘工作中，许多新的有价值的模式将被发现。同时，基于分布式数据挖掘的算法研究将会有越来越多的研究人员投入其中，以实现对不同数据源、多种数据库间的数据挖掘。

（4）专门用于知识发现的数据挖掘形式化[34]的描述语言 DMQL 将走向形式化和标准化。数据挖掘领域的一般研究方法将越来越统一。同时也会被运用到森林生态站的数据挖掘中，建立标准化的、统一的数据分析平台。

（5）可视化数据挖掘过程[35]，能够使得数据的挖掘结果很容易的被使用者理解。提高人与机器的交互能力，使数据挖掘成为使用者业务流程的一部分。这也将降低森林生态站数据的使用标准，普通人也可以通过良好的人机交互程序从海量的数据中获取自己感兴趣的内容。

（6）通过数据挖掘所发现的各种模式将会成为系统服务的一部分，面向普通用户公布，通过这个途径可以将数据转化为市场价值。最终通过市场的需求推动森林生态站数据挖掘研究的良性发展。

（7）新的数据类型的数据挖掘需求。通过建立新的森林生态站数据存储格式将会产生基于新的数据类型的数据挖掘技术。处理数据类型的复杂性，就需要新的分析和建立模型的方法。就也将成为未来研究的重点方向之一。

（8）对知识的维护更新。通过新的数据挖掘方法的引入会有越来越多的模式发现。基于已有的模式进行下一步挖掘工作，也将成为未来数据挖掘的新的研究方面。

（9）实时数据挖掘[36]。森林生态站的数据近似于实时采集（时间间隔短），就目前的情况看数据的分析工作相对于数据的产生时间严重滞后。数据经设备采集之后通常是缓存在采集设备中，由数据采集人员定期人工导出。数据时效性的价值被严重降低，建立数据的实时传回机制，进行实时的数据挖掘分析，也将成为未来研究的重点任务之一。

（10）预测性数据挖掘。基于几十年森林生态站积累的观测数据，进行预测性数据挖掘也将具有十分重要的意义。通过分析，研究人员可以发现自然的变化规律，以及自然生态体系健康状况，以此来指导生活。

参考文献

[1] 冯振兴，陈志泊. 数字化森林生态站的构建研究[J]. 农业网络信息. 2007(11)

[2] 沈静，陈志泊，王兵，李少宁. 数字化森林生态站系统的构建[J]. 农业网络信息. 2005(12)

[3] 王兵，李少宁. 数字化森林生态站构建技术研究[J]. 林业科学. 2006(01)

[4] 王树良，丁刚毅，钟鸣. 大数据下的空间数据挖掘思考[J]. 中国电子科学研究院学报. 2013(01)

[5] 丁岩，杨庆平，钱煜明. 基于云计算的数据挖掘平台架构及其关键技术研究[J]. 中兴通讯技术. 2013(01)

[6] 董肖莉. 森林生态站时序数据的模式挖掘系统研建[D]. 北京林业大学, 2012.

[7] 程陈 . 大数据挖掘分析 [J]. 软件，2014，35(4):130-131

[8] 王新华，米飞，冯英春，赵玮. 空间数据挖掘技术的研究现状与发展趋势[J]. 计算机应用研究. 2009(07)

[9] 张慧萍，陈志泊. 分布式数据库技术在数字化森林生态站的应用[J]. 北京林业大学学报. 2007，29（3）：131-135

[10] 刘威. 分布式数据库及其技术[J]. 长春大学学报. 2000(01)

[11] 靳皞. 浅谈分布式数据库系统的设计与优化[J]. 硅谷. 2011(02)

[12] 林源， 陈志泊. 分布式异构数据库同步系统的研究与应用[J]. 计算机工程与设计， 2010， (24)

[13] 刘鹏，雷蕾，张雪凤. 缺失数据处理方法的比较研究[J]. 计算机科学. 2004(10)

[14] 张靖，姚珍，唐雪飞. 基于决策树的不完整数据的处理[J]. 电子科技大学学报. 2007(01)

[15] 黄创光，印鉴，汪静，刘玉葆，王甲海. 不确定近邻的协同过滤推荐算法[J]. 计算机学报. 2010(08)

[16] 李长军. 基于贝叶斯网络的中医医案数据挖掘[D]. 厦门大学 2008

[17] 廖学清. 数据缺失下学习贝叶斯网的研究[D]. 苏州大学 2008

[18] 朱金清， 王建新， 陈志泊. 基于 APRIORI 的层次化聚类算法及其在 IDS 日志分析中的应用[J]. 计算机研究与发展 ISTIC EI PKU， 2007， 44

[19] 金连，王宏志，黄沈滨等. 基于 Map-Reduce 的大数据缺失值填充算法[J]. 计算机研究与发展 ISTIC EI PKU，2013， 50

[20] 徐胜利. 一种堆栈式快速等值线图填充算法[J]. 计算机工程与应用， 2010， (8)

[21] 杨涛， 骆嘉伟， 王艳等. 基于马氏距离的缺失值填充算法[J]. 计算机应用， 2005， (12).

[22] 田秀华， 王忠宝， 张展. 基于连续傅里叶变换计算离散傅里叶变换的一种算法[J]. 自动化技术与应用，2007， 26(8). DOI:10.3969/j.issn.1003-7241.2007.08.013.

[23] 张宪超， 武继刚， 蒋增荣等. 离散傅里叶变换的算术傅里叶变换算法[J]. 电子学报， 2000， (5)

[24] 张师超， 朱曼龙， 黄樑昌. QENNI:一种缺失值填充的新方法[J]. 广西师范大学学报(自然科学版)， 2010，(1)

[25] Lee W，Stolfo S J，Chan P K，et al.Real Time Data Mining-based Intrusion Detection. Proc Second DARPA Information Survivability Conference and Exposition . 2001

[26] Jonathan L. Herlocker，Joseph A. Konstan，Loren G. Terveen，John T. Riedl. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems (TOIS) . 2004 (1)

[27] Xiaoyuan Su，Taghi M. Khoshgoftaar，Jun Hong. A Survey of Collaborative Filtering Techniques[J]. Advances in Artificial Intelligence . 2009

[28] Liu Zhao Jiang Liangxiao(Institute of Information and Engineering of China University of Geosciences，Wuhan430074).The Research of Data Mining Based on Neural Networks. Computers in Engineering . 2004

[29] Akhil Kumar. New Techniques for Data Reduction in a Database System for Knowledge Discovery Applications[J]. Journal of Intelligent Information Systems . 1998 (1)

[30] Bingru Y， Jiangtao S. KDD ＊ based on double-base cooperating mechanism and its realization of software[J]. Systems Engineering and Electronics， Journal of， 1999， 10(4):1 – 9

[31] D. Pyle.Data Preparation for Data Mining. . 1999

[32] Tavani H T. KDD， data mining， and the challenge for normative privacy[J]. Ethics and Information Technology， 1999， 1(4):265-273

[33] Han J.Conference tutorial notes : data mining technique. Proceedings of the 1996 ACM-SIGMOD International Conference On Management of Data (SIGMOD’96) . 1996

[34] Shin C， Yun U T， Kim H K， et al. A hybrid approach of neural network and memory-based learning to data mining[J]. Neural Networks， IEEE Transactions on， 2000， 11(3):637 – 646

[35] Kiem H， Phuc D. New Directions in Rough Sets， Data Mining， and Granular-Soft Computing[M]. Springer Berlin Heidelberg， 1999:448-452.

[36] Liu Jingxue & Fei Qi 1. Dept. of Information Warfare， Wuhan Commanding Communications Academy，Wuhan 430010， P. R. China; 2. Inst. of Systems Engineering， Huazhong Univ. of Science and Technology，Wuhan 430074， P. R. China.. Research of intelligence data mining based on commanding decision-making[J]. Journal of Systems Engineering and Electronics. 2007(02)

二、研究方案

1. 研究内容、研究目标及拟解决的关键问题

1. 研究的内容

(1) 基于皮尔逊积矩相关系数（Pearson product-moment correlation coefficient 简称：PPMCC）的观测指标相关性研究。通过 PPMCC 算法，能够得到观测指标之间的相关性，为数据填充算法选取样本数据的理论依据。

(2) 加权的基于象限最近邻填充算法（Weighted Quadrant Encapsidated Nearest Neighbor based Imputation 简称：WQENNI）的数据补全研究。WQENNI 算法由象限最近邻填充算法（Quadrant Encapsidated Nearest Neighbor based Imputation 简称：QENNI）改进而来。在实际的数据挖掘工作中，由于各种人力不可控的因素，产生数据缺失是非常普遍的，这将严重影响到从数据中发现的模式的正确性和准确性，因此，采用一定的方法对缺失的数据进行补全是一个不可或缺的工作。不同于已被大家熟知的 k 最近邻填充（k-Nearest Neighbor algorithm 简称：kNN）算法，虽然 kNN 是一种优秀的填充算法，但是 kNN 算法选择的缺失数据的最近邻可能偏向某一边，这样将导致极大的偏差。此外，k 值的设定将极大程度上影响算法的正确性和准确性。在实际的挖掘过程中，如果 k 值过大，随机性太严重；k 值过小，达不到统计学上的大样本容量标准。采用 QENNI 算法将尽可能地消除对 k 值的依赖，该算法对于低维数据集可以使无参。在现有的 QENNI 算法基础上，对于各象限上找到的最近邻进行加权，然后计算缺失的数据。实验表明 QENNI 的算法的填充准确性要优于 kNN 算法。

(3) 基于离散时间傅里叶变换的聚类算法（Discrete-time Fourier Transform Clustering Algorithm 简称：DTFTCA）森林生态站观测指标聚类分析。傅里叶变换是将信号在时域（或空域）和频域之间变换时使用。对信号进行时域分析时，有时一些信号的时域参数相同，但并不能说明信号就完全相同。因为信号不仅随时间变化，还与频率、相位等信息有关，这就需要进一步分析信号的频率结构，并在频率域中对信号进行描述。傅立叶定理使得以上的分析成为可能，该定理表明任何连续测量的时序或信号，都可以表示为不同频率的正弦波信号的无限叠加。数学家傅立叶在 1822 年证明了这个著名的定理，并创造了为大家熟知的、被称之为傅立叶变换的算法，该算法利用直接测量到的原始信号，以累加方式来计算不同正弦波信号的频率、振幅和相位。森林生态站指标的观测数据都是基于时序的数据，很难直接在时域上通过距离公式发现其之间潜在的相似周期性变化。通过傅里叶变换将其转换为频域上的变化，然后利用距离公式计算观测指标之间的距离，根据距离的远近判断指标之间是否具有相同的周期性变化，以此为依据将观测指标聚类。

(4) 森林生态站数据管理平台（Forest Ecology Station Data Management Platform 简称：FESDMP）统一的生态站数据管理平台主要是实现森林生态站观测指标管理为目的的，以数据的预处理，导入、导出和数据挖掘和分析为主要功能的综合数据管理平台。系统在设计的过程中需要考虑海量数据的存储，以及完善的权限设置功能和可伸缩的数据挖掘处理工作的实现。

2. 研究的目标

(1) 将加权后的基于象限近邻填充算法（WQENNI）的数据补全技术应用到森林生态站观测指标的观测数据补全上。

(2) 通过该算法对数据进行预处理后，采用离散时间傅里叶变换的聚类算法（DTFTCA）将基于时序的数据转化为频域上的数据，之后采用距离公式对观测指标进行聚类分析。

(3) 最终将该算法运用到生态站数据管理平台中，提供对外的观测指标聚类分析服务。

3. 拟解决的关键问题

针对课题研究内容，探讨课题拟解决的关键问题如下：

(1) 基于皮尔逊积矩相关系数（Pearson product-moment correlation coefficient 简称：PPMCC）的观测指标相关性研究。通过 PPMCC 算法，能够得到观测指标之间的相关性，为数据填充算法选取样本数据的理论依据。

(2) 加权的基于象限近邻填充算法（WQENNI）的数据补全技术在森林生态站数据处理工作中的应用。数据挖掘的前期工作数据预处理和缺失值补全是保证数据挖掘算法发现的模式的正确性和准确性的先决条件。但是算法的直接迁移并不能直接运用到森林生态站的特定类型的数据补全上，针对该问题，需将该算法进行改进。对于缺失值的填充，填充的准确性是十分重要。本系统也将用 WQENNI 与 kNN 算法的补全结果进行对比。

(3) 基于离散时间傅里叶变换的聚类算法（Discrete-time Fourier Transform Clustering Algorithm 简称：DTFTCA）的实现。傅里叶变换的运用已经十分广泛，但是将其运用到森林生态站的数据聚类分析上，在国内外本课题是首创。同时，结合我们所拥有的数据的特点，将时域上的数据转化为频域上的变换是一次大胆的尝试。在频域上通过距离公式求出各个指标之间的距离，拟通过基于密度的聚类算法（Density-based Spatial clustering of applications with noise 简称 DBSCAN）对森林生态站的观测指标进行聚类分析。

(4) 生态站数据管理平台的搭建。结合数据量和数据特点等因素，平台搭建需要满足一下基本要求。海量数据存储方案。单一的数据库存储已经不能满足森林生态的数据存储需求，分布式存储方案将被引入到该平台。系统拟采用 SQL 数据库存储，为了避免单表数据量大，影响查询速度，系统将自动滴为每个指标每年生成一张存储表。这样将有利于数据的存储于查询。该平台是基

于 B/S 架构，为了保证系统的交互性，前端设计上将采用 ExtJS 4.x 来开发基于 Web 的 RIA 程序。

2. 拟采用的研究方法、技术路线、实验方案及可行性分析

1. 拟采用的研究方法

(1) 通过阅读相关文献了解缺失数据补全、聚类算法等领域应用研究现状，确定课题及课题研究方法。通过学习当前大数据分析方案，确立森林生态站数据管理平台的架构设计。

(2) 数据的获取与预处理

本课题需要大量的森林生态站观测指标的观测数据，拟通过与水土保持学院合作开发森林生态站数据管理平台获得。

(3) 算法分析

- 拟采用改进的 QENNI 算法进行森林生态站数据的缺失补全，重点工作在于将改进后适用于森林生态站数据的算法应用到数据预处理工作中。同时，将把 WQENNI 的补全准确率与 kNN 算法进行对比分析。

数据样本的选取 数据样本拟采用两种方案进行对比，最终根据实验结果选取填充率最高的方案。

1. 选取不同年份的相同时间段的指标的离散观测值。比如缺失 2014 年 7 月 1 日 12:00 的观测值，则选取 2004 年~2014 年的 7 月 1 日观察的观测值作为相关数据。
2. 依据皮尔逊积矩相关系数相关性分析，选取 n 个具有相关性的观测指标，获取该 n 个指标 2014 年 7 月 1 日的所有观测数据，作为 QENNI 算法的输入数据。

皮尔逊积矩相关系数 用于度量两个变量 X 和 Y 之间的相关（线性相关），其值介于-1 与 1 之间。在自然科学领域中，该系数广泛用于度量两个变量之间的相关程度。

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

公式一

QENNI 算法过程 假如 X 是一个 m 维随机向量，Y 是受 X 所影响的因变量。实际中，如果取得的一个含缺失数据的随机样本（样本大小为 n），可以表示为（X_i，Y_i，δi），i=1，2，…，n。其中所有的 X_i 向量是可观测的，当 Y_i 缺失时，记 δ_i=1，否则记 δ_i=0。如果数据集 T 含有 n 个数据，每个数据有 m+1 个数据，每个数据有 m+1 个属性（包括 m 个条件属性和 1 个决策属性），记为：Ti=（X_{i1}，X_{i2}，…，X_{im}，Y）（文本的缺失值仅在决策属性 Y 中产生。）T=I ∪C，让 r=$\sum_{i=1}^{n}$δi，其

中 I={T_1，…，Tr}，r≤n 是决策属性缺失的数据集，简称缺失数据集；C={T_{r+1}，…，T_n}是完全数据集。

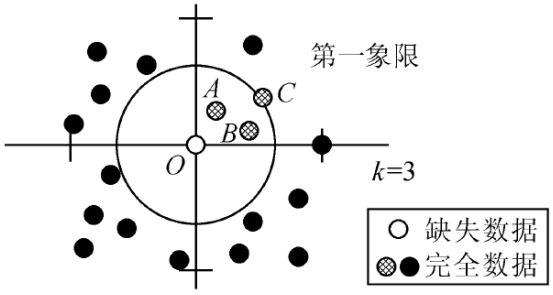

图 1 kNN 算法最近邻点的选择

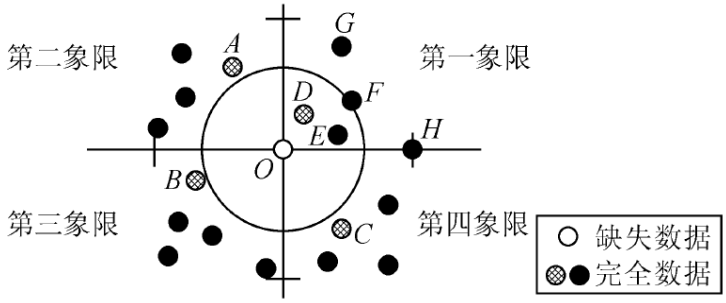

图 2 QENNI 算法最近邻点的选择

kNN 算法选取的 k 个最近邻点可能会有偏好，这使得填充效果相对低效。从实际上看，缺失数据周围的完全数据从分布情况上看，其最近邻数据和总体数据的分布可能是不一致的。按图 1 所示，O 点代表缺失数据，其他点代表完全数据，我们用 kNN 来填充缺失数据，找到的 k=3 的三个临近点就是 A、B、C，按照这种方式计算的缺失值偏向于第一象限。这不是我们想要看到的结果。

kNN 算法的最佳 k 选取比较困难，对每个数据集进行 kNN 实验前，必须反复计算得到最优的 k。同时，k 一旦偏差过大，将大大影响算法的填充性能。因此，如果算法能够消除对 k 的依赖就会成为最佳选择。

分析发现，用来估计缺失值的完全数据必须在近邻的情况下，尽可能地在缺失数据的第一个环形圆上，而且算法要最大限度地消除对参数 k 的依赖。

基于象限最近邻填充算法（Quadrant Encapsidated Nearest Neighbor based Imputation 简称：QENNI）。首先，把包含 m 个条件属性（X_1，X_2，…，X_m）的数看做 m 维空间上的一个点。并对该 m 维空间进行构造，以缺失数据为中心建立坐标系，通过坐标轴把 m 维空间划分成 2^m 象限。

当 m=2，数据的条件属性可以看做平面上的一个点。如图 2 所示。当某点恰好处于两象限间，把该点归到它的某个近邻象限，根据构造，任意数据集的一数据，都会位于唯一确定的象限中。

同理，当 m= 3 时，数据的条件属性为（X_1，X_2，X_3），形成的空间坐标系把空间分成 2^3 个象限。每个数据也都位于唯一确定的象限中。

类似的，拓展到一般情况。对于一般的 m，数据的条件属性为（X_1，X_2，…，X_m）以缺失数据为中心形成的坐标系可把 m 维空间分成 2^m 个象限。同样，把处于象限间的临界数据归到它的某个邻近象限。显然，数据集中任一数据也都位于唯一确定的象限中。

基于上面对 m 维空间的划分，在每个象限中找到一个离缺失数据最近的完全数据（若象限中不存在完全数据则该象限不取），并将这些选中的完全数据用于填充缺失值。以 m= 2 为例，如图 2，把平面分成 4 个象限，在第一象限离缺失数据 O 距离最近的是点 D，于是该象限只选用该点，相应的在第二、三、四象限分别只取点 A，点 B 和点 C，然后用已选出的点 A、B、C 和 D 的决策属性来加权填充点 O 的缺失决策属性。

综上对 QENNI 算法的过程，可以做出如下定义：

定义 1 以 T_i 为中心的坐标系把 m 维空间分成 2^m 个象限,完全数据集 C 基于象限也被分成 2^m 个子集 C= {D_1，D_2，…，D_q，…，D_2m}，每个完全数据集 D_q(q= 1，2，…，2^m) 称为 T_i 第 q 象限的数据。

比如，如图 2 所示，以 To 为中心的直角坐标系把平面分成 4 个象限，对于第一象限，D_1= { T_D，T_E，T_F，T_G，T_H } (临界点 T_H 也归到该象限中) 就是 To 第一象限的数据。

定义 2 $\forall Tj \in D_q$，满足 $Near_q = \arg\min dist(T_i, T_j)$，$T_j \in D_q$ 则称 $Near_q$ 为 T_i 第 q 象限的最近邻点。比如，如图 2 所示，以第一象限为例，T_o 第一象限的数据 D1= { T_D，T_E，T_F，T_G，T_H }，它离缺失数据 T_o 距离最近的数据为 T_D，$Near_q$= T_D 就是 T_o 第一象限的最近邻点。

定义 3 在第 q 象限范围内，以 T_i 为(超) 球中心，dist ($Near_q$，T_i) 为半径所确定的(超) 球面，称为 T_i 第 q 象限的壳 $Shell_q$。

定义 4 由所有 $Shell_q$(q= 1，2，…，2_m)，以及坐标轴构成的(超) 平面一起围成的 m 维子空间，称为 T_i 的壳 Shell。

定义 5 把 T_i 在每个象限中的最近邻点{Near1，Near2，…，Near2^m } 统称为壳层的点。

根据 QENNI 算法和上述定义，可以得到以下性质：

性质 1 若 $D_q \neq \emptyset$，则第 q 象限必存在 T_i 该象限的壳 $Shell_q$。

性质 2 $\forall Tj \in D_q$，$\exists$dist(Tj，Ti) $\geq$ dist (Nearq，Ti)

为克服 kNN 在 k 个最近邻选择上可能有偏好的问题(如图 1)，在存在的情况下，选出的近邻点应尽可能满足：a.选出的最近邻完全数据不会偏向缺失数据的一边，即要把缺失数据围住；b.选出的

完全数据把缺失数据围住的壳是最小的，即缺失数据的壳中必须不存在其他完全数据。事实上，只要存在完全数据能围住缺失数据，QENNI 算法找出的缺失数据的壳，是可以满足以上两点的。

QENNI 算法显然对 a 成立， 下面我们来证明一下它也满足 b。可把 b 等价描述为：已知缺失数据 T_i，Shell 是 T_i 的壳。证明 T_i 的壳 Shell 里面不存在其他完全数据。

证明：假设缺失数据的壳 Shell 里面还有其他完全数据，不妨设壳里面还存在一个完全数据 T_j（$Tj \in C$），由性质 1，进一步假设 T_j 就是缺失数据第 q 象限的壳 $Shell_q$ 里的完全数据，并记 $Near_q$ 为 T_i 第 q 象限的最近邻点。因为 T_j 在第 q 象限的壳里面，可得到 dist (T_j，T_i) < dist ($Near_q$，T_i)。而由性质 2，dist (T_j，T_i) ≥dist($Near_q$，T_i)，显然两式是矛盾的。因此假设错误。T_i 的壳 Shell 里面不存在其他完全数据。

于是，缺失数据的壳中不再存在其他完全数据，即所得到的缺失数据的壳是最小的。可见，QENNI 算法找出的完全数据在满足不偏好的前提下是离缺失数据最近的，它们最能代表缺失数据。

算法改进

通过选出的偏好点的决策数据来计算缺失值，该算法未对加权方法进行说明。如果直接利用计算出的决策数据简单平均，将严重影响算法的准确性。因此，明显的改进是根据计算得到的邻近点的贡献加权。将较大的权值赋给较近的近邻。

可用距离平方的倒数作权值：

$$\hat{f}(x_q) \leftarrow \underset{v \in V}{\operatorname{argmax}} \sum_{i=1}^{k} w_i \delta(v, f(x_i))$$

其中：

$$w_i \equiv \frac{1}{d(x_q, x_i)^2}$$

公式二

方法存在问题，恰好匹配当前样例点会将导致分母趋无穷的。如果出现此情况作特殊处理。

研究发现，数据可能偏向于分布在某一个象限空间。

算法的下一步研究点 是方向距离综合加权，降低数据过多的分布在某几个象限而仅仅选取一个值对数据的准确性的影响。

- 基于离散时间的傅里叶变换的聚类算法（ DTFTCA）的实现。

傅里叶变换是一种线性积分的变换，常用在将信号在时域（或空域）和频域之间变换，在物理学和工程学中有许多应用。傅里叶变换源自对傅里叶级数的研究。傅里叶级数指的是任何周期函数都可以用正弦函数和余弦函数构成的无穷级数来表示（选择正弦函数与余弦函数作为基函数是因为它们是正交的）。

连续傅里叶变换 一般情况下，若“傅里叶变换”一词不加任何限定语，则指的是“连续傅里叶变换”（连续函数的傅里叶变换）。连续傅里叶变换将平方可积的函数 $f(t)$表示成复指数函数的积分或级数形式。

$$F(\omega) = \mathcal{F}[f(t)] = \int_{-\infty}^{\infty} f(t) e^{-i\omega t}\,dt.$$

公式三

公式二是将频率域的函数 F(ω)表示为时间域的函数 $f(t)$的积分形式。

连续傅里叶变换的逆变换（Inverse Fourier transform）

$$f(t) = \mathcal{F}^{-1}[F(\omega)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t}\,d\omega.$$

公式四

傅里叶级数 连续形式的傅里叶变换其实是傅里叶级数（Fourier series）的推广，因为积分其实是一种极限形式的求和算子而已。

$$f(x) = \sum_{n=-\infty}^{\infty} F_n\, e^{inx},$$

公式五

其中 F_n 为复振幅。对于实值函数，函数的傅里叶级数可以写成：

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left[a_n \cos(nx) + b_n \sin(nx)\right]$$

公式六

离散时间傅里叶变换 离散时间傅里叶变换（DTFT，Discrete-time Fourier Transform）以离散时间 nT (其中n ∈ Z，T 为采样间隔) 作为变量函数（离散时间信号）f(nT)变换到连续的频域，即产生这个理算时间信息的连续频谱 $F(e^{i\omega})$，值得注意的是这一频谱是周期的。

记连续时间信号 f(t)的采样为：

$$f_{sp}(t) = \sum_{n=-\infty}^{\infty} f(t)\delta(t - nT)$$

公式七

其傅里叶变换为：

$$\mathfrak{F}\{f_{sp}(t)\} = \int_{-\infty}^{\infty} f_{sp}(t) e^{-i\omega t}\,dt = \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(t)\delta(t-nT) e^{-i\omega t}\,dt = \sum_{n=-\infty}^{\infty} f(nT)\, e^{-in\omega T}$$

公式八

这就是采样序列 f(nT)的 DTFT：

$$F_{DTFT}(e^{i\omega T}) = \sum_{n=-\infty}^{\infty} f(nT)\, e^{-in\omega T}$$

公式九

为方便起见，通常将采样间隔 T 归一化，即为 f(nT)的离散时间傅里叶变换：

$$F_{DTFT}(e^{i\omega}) = \sum_{n=-\infty}^{\infty} f(n)\, e^{-in\omega}$$

公式十

其反变换为：

$$f(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F_{DTFT}(e^{i\omega})\, e^{in\omega}\, d\omega$$

公式十一

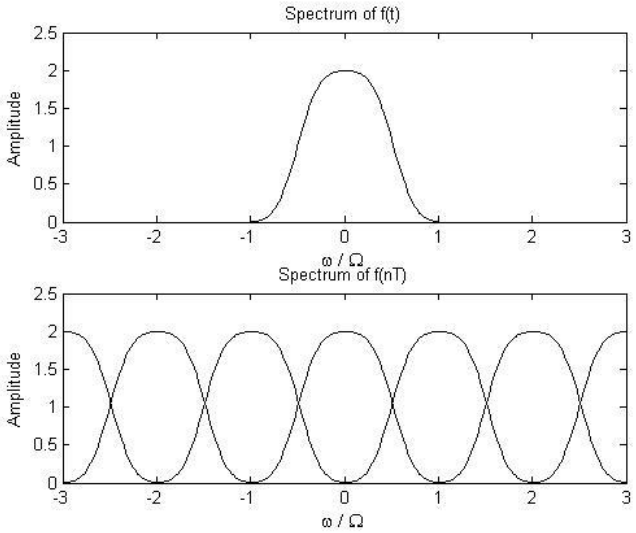

图 3 周期性离散时域到频域变换对比

基于密度聚类的算法 DBSCAN 介绍 DBSCAN 是典型的基于密度的聚类算法。与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。

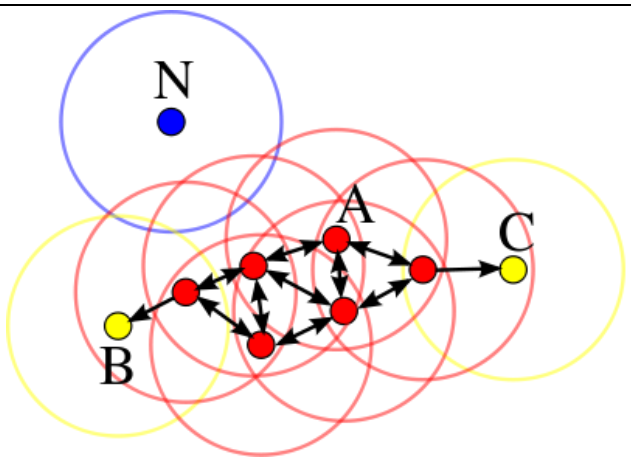

图 4 红色为核心店，黄色为边界点，蓝色为噪声点

DBSCAN 算法中两个重要参数：Eps（定义密度时的领域半径）和 MinPts（定义核心点时的阈值），分别简记为 e 和 M。考虑数据集合 $X=\{X_1,\ldots,X_n\}$，首先介绍一下概念与记号。

e 邻域（e neighborhood）：

设 $x\in X$，称

$$N_e(x)=\{y\in X:d(y,x)\le e\}$$

为 x 的 e 邻域。显然，$x\in N_e(x)$。

密度（density）：

设 $x\in X$，称

$$\rho(x)=|Ne(x)|$$

为 x 的密度。注意，这里的密度是一个整数值，且依赖于半径 e。

核心点（core point）：

设 $x\in X$，若 $\rho(x)\ge M$，则称 x 为 X 的核心点。记由 X 中所有核心点构成的集合为 X_c，并记 $X_{nc}=X\setminus X_c$ 表示由 X 中的所有核心点构成的集合。

边界点（border point）：

若 $x\in X_{nc}$，且 $\exists y\in X$，满足

$$y\in N_e(x)\cap X_c$$

即 x 的 e 邻域中存在核心点，则称 x 为 X 的边界点。记由 X 中所有的边界点构成的集合为 X_{bd}。

噪音点（noise point）：

记 $X_{noi}=X\setminus(X_c\cup X_{bd})$，若 $x\in X_{noi}$，则称 x 为噪音点。且满足 $X=X_c\cup X_{bd}\cup X_{noi}$，如图 4 所示。

直接密度可达（directly density reachable）：

设 $x,y\in X$，若满足 $x\in X_c$，$y\in N_e(x)$，则称 y 是从 x 直接密度可达的。

密度可达（density reachable）：

设 $P^{(1)},P^{(2)},\cdots,P^{(m)}$，其中 $m\geq 2$ 。若他们满足：$P^{(i+1)}$是从 $P^{(i)}$直接密度可达的，$i=1,2,...,m-1$，则称 $P^{(m)}$是从 $P^{(1)}$密度可达。

密度相连（density connected）：

设 $x,y,z\in X$，若 y 和 z 是从 x 的密度可达的，则称 y 和 z 是密度相连的。显然，密度相连是具有对称性的。

类（cluster）：

称非空集合 $C\subset X$ 是 X 的一个类（cluster），如果它满足：对于 $x,y\subset X$

（1）（Maximality）若 $x\in C$，且 y 是从 x 密度可达的，则 $y\in C$。

（2）（Connectivity）若 $x\in C,y\in C$，则 x,y 是密度相连的。

算法核心思想：

从某个选定的核心点出发，不断向密度可达的区域扩张，从而得到一个包含核心点和边界点的最大化区域，区域中任意两点密度相连。

考虑数据集合 X，DBSCAN 算法的目标是将数据集合 X 分成 K 个 cluster（注意 K 也由算法得到，无需事先指定）及噪音点集合，为此，引入 cluster 标记数组

$$m_i=\begin{cases} j\ (j>0), & \text{若 } x^{(i)} \text{ 属于第 } j \text{ 个 cluster;}\\ -1, & \text{若 } x^{(i)} \text{ 为噪音点,}\end{cases}$$

由此，DBSCAN 算法的目标就是生成标记数组 $m_i,i=1,2,...,N$，而 K 即为 $\{m_i\}_{i=1}^{N}$ 中互异的非负数的个数。

基于离散时间傅里叶变换的聚类算法设计

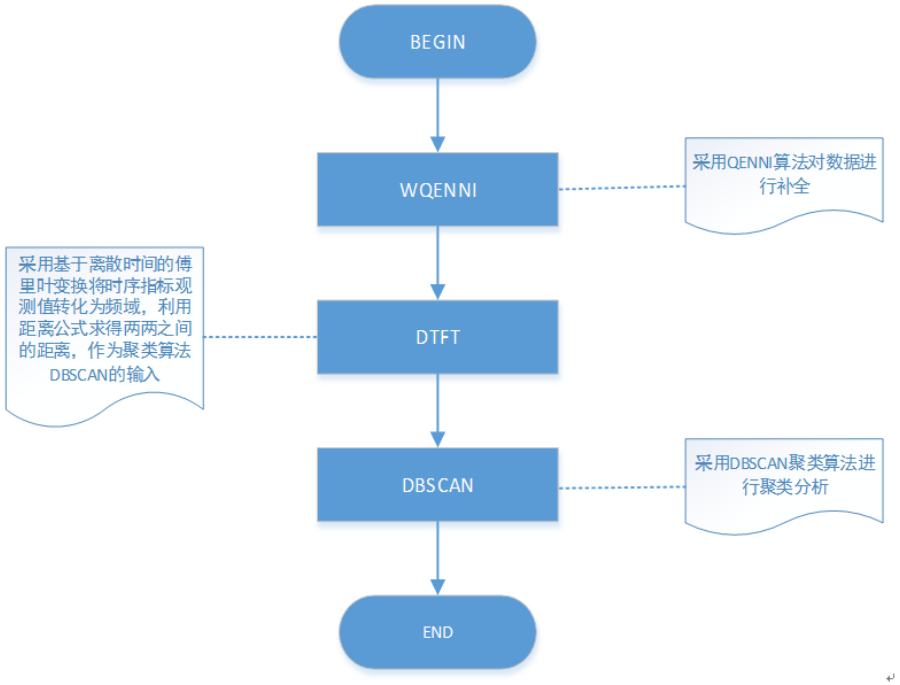

图 5 聚类分析算法整体流程图

综上对基于离散时间傅里叶变换的聚类算法（Discrete-time Fourier Transform Clustering Algorithm 简称 DTFTCA）定义如下：

定义 1 定义集合 F={f1(t)，f2(t)，…，fk(t)}为观测指标在连续时间上的采样。

定义 2 依据***公式九***得基于集合 F 的 f(nT)的 DTFT 集合为 E={FDTFT1，FDTFT2，…，FDTFTk}。

定义 3 依据***公式十***将采样间隔时间 T 归一化得集合 G={F1(n)，F2(n)，…，Fk(n)}。

定义 4 遍历集合 G，采用积分的方式算出任何两条线之间的距离，得到集合 D={ D_{ij}}，且 $1 \le i \le k, 1 \le j \le k$ 。

$$d(x,y) := \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

定义 5 根据 D 集合采用算法 DBSCAN 算法，进行聚类分析。

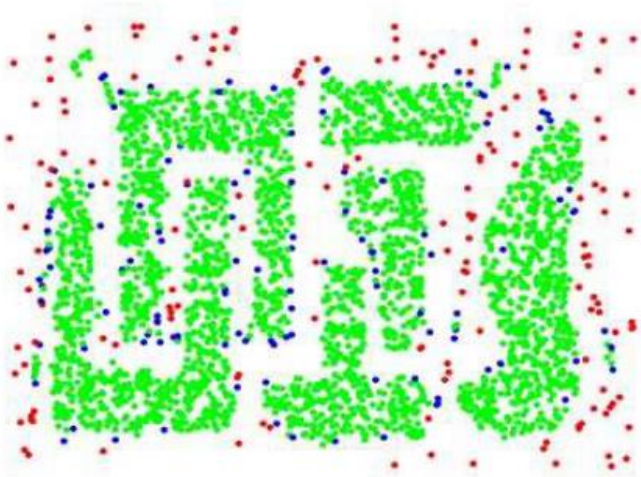

图 6 聚类分析结果图

算法验证

基于离散傅里叶变化的密度聚类算法在理论上已设计实现，后续将用相同的数据集使它与其他聚类算法对比，然后进一步改进该聚类算法。

（4）平台建设

通过调研当前大公司大数据的存储方案以及基于浏览器的 RIA 企业级程序开发架构，确定本平台的系统建设方案，并积极投入到实施中。

2. 拟采用的技术路线与实验方案

考虑到基于 Java 的平台，很多数据挖掘研究人员已经经过大量的实践，积累了丰富的数据挖掘工具和学习的文档，这将给我的研究带来极大的便利。同时，基于 Java 的大数据处理平台 Hadoop 也是十分成熟的方案。为了程序的可扩性和可迁移性，以及对 Hadoop 平台的兼容性，本课题也将在 Java 平台下实现相关的算法与森林生态站数据管理平台搭建工作。

QENNI 算法主要实现森林生态站观测数据的填充任务，傅里叶变换实现聚类分析。

3. 可行性分析

数据挖掘是一个动态的、强势快速扩展的领域。经过这么多年的发展，该领域已经成熟。特别是相关的基础理论、挖掘方法的不断完善，以及其在商业应用、Web 数据挖掘、电子商务等领域的典型成功应用更是极大地促进了数据挖掘科学的发展。

在理论方面，通过阅读大量相关文献，已经了解到数据挖掘应用的热门研究策略，并在现有的数据填充策略基础上进行了优化研究制定了新的研究方法，并针对森林生态站基于时序的数据引入傅里叶变换聚类方法。

技术方面，算法程序采用 Java 编程语言。本人有近四年的 Java 编程开发经验，相信在算法的

实现上没有阻碍。

项目建设上，本人有相对丰富的项目开发、管理经验。从事过北林数字标本馆、成人教育学院学籍管理系统等众多项目的开发。相信在工程实施上能在保证质量的前提下，按时完成。

3、本研究的特色与创新之处

数据挖掘技术虽然早已在众多行业展现其强大的使用价值，但是在森林生态站数据分析领域还处于发展阶段。基于 QENNI 算法改进的缺失数据填充算法更是优于传统的 kNN 算法。基于傅里叶变换密度聚类算法用于森林生态站观测指标聚类分析也是其创新之处。国内外的文献资料表明，利用傅里叶变换来对数据进行聚类分析在研究领域还没有研究人员这样实验过。

本课题的另一个特色之处是统一的森林生态站数据管理平台的建立。纵观国内的森林生态站数据管理现状，平台一旦建立将给该领域的研究人员和管理人员带来极大的便利。

4、研究计划及预期研究结果

1. 研究计划

2014.03 - 2014.06 阅读文献，查阅资料，选题

2014.07 – 2014.09 算法设计，编写程序，，撰写文献综述及开题报告

2014.10 - 2015.08 优化程序代码，撰写小论文及，完成毕业论文初稿

2015.09 - 2016.04 修改毕业论文，定稿，准备答辩

2. 预期成果

1. 优化并实现基于象限近邻填充算法在多维空间上的数据补全
2. 优化并实现基于傅里叶变换密度聚类算法在森林生态站观测数据上的使用
3. 发表 EI 或核心期刊论文一篇

三、研究基础

1、已参加过的相关研究工作和已取得的研究工作进展

1.1 相关研究工作

（1）本科期间参加过国家级大学生创新项目。通过该项目，科研能力和方法都得到了一定程度的锻炼。

（2）在研究生阶段参与过多个项目的开发任务。取得三个软件著作权。

（3）通过阅读相关文件和导师的知道，对该领域结束有一定的了解和掌握，提出的技术算法方案有创新性并切实可行。

1.2 相关研究工作

（1）算法的程序代码已实现 30%，基于 QENNI 的缺失数据填充算法改进的建模工作，基于时域的求解两个指标的距离，以及采用傅里叶变换后的求解两个指标的距离的基于密度聚类的相关算法建模工作。

（2）森林生态站数据管理平台已完成需求分析、概要设计，目前正在紧密的开发之中。项目初步预估完成 30%。

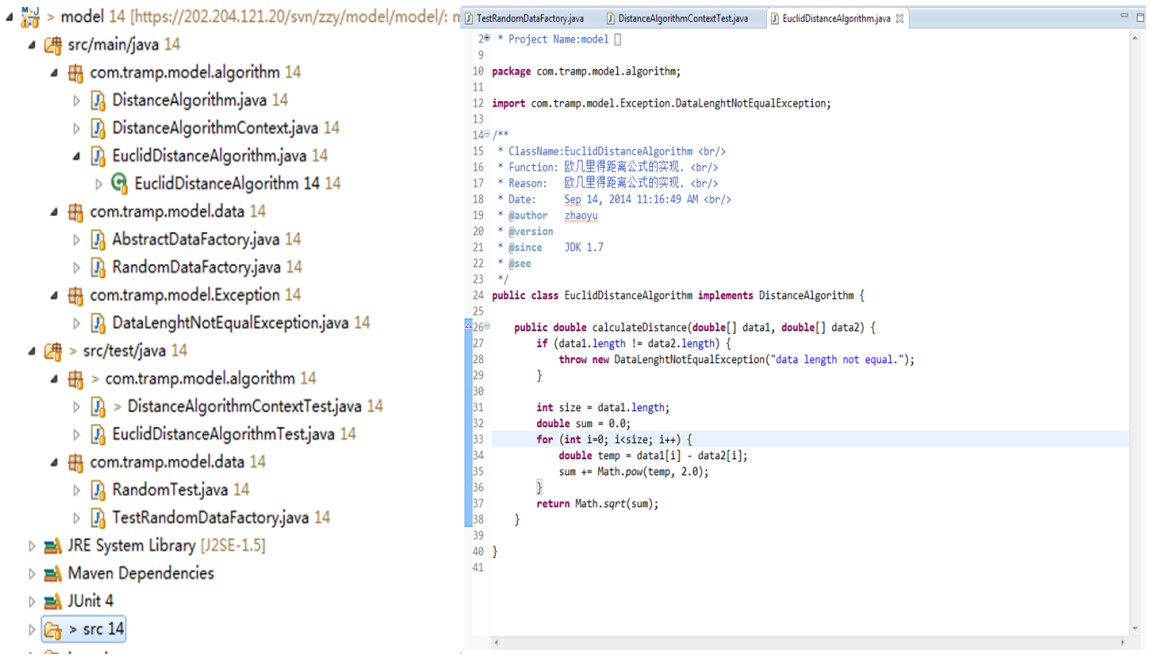

图 7 已完成的算法代码

图 8 森林生态站数据管理平台界面

2、已具备的实验条件，尚缺少的实验条件和拟解决的途径

实验室硬件设施齐全，算法的代码已在完善中。森林生态站的数据管理平台也紧密的开发之中。目前上缺少的实验条件是真实的森林生态站的观测指标的观测数据。针对该问题，在实验阶段会随机的模拟一些数据，在数据管理平台开发完成，会从水保学院获取真实的数据导入管理系统，届时可以采用真实的数据对算法进行验证。

3、研究经费预算和经费落实情况

研究经费集中在论文发表方面和森林生态站数据管理平台的开发以及平台运行的硬件服务器上，有一定经费支持。

四、导师对开题报告的评价

（就硕士生对国外研究现状的了解情况、研究内容、研究方法、预期成果等方面予以评价）

导师签字：

年　　月　　日

五、开题报告小组评议意见

组成	姓名	职称	工作单位	本人签字
组长				
成员				

开题报告小组评议意见（如选择项请划√）

(1) 论文选题有无理论和实践意义	□ 选题具有很强的理论意义和实用价值 □ 选题具有较强的理论意义和实用价值 □ 选题缺乏理论意义和实用价值
(2) 文献阅读是否全面反映与研究课题相关的现状和发展趋势	□ 文献综述全面阐述该研究方向的现状和发展动态 □ 文献综述基本跟踪该研究方向的现状和发展动态 □ 综述一般，未达到上述标准
(3) 研究方案是否可行	□ 可行 □ 基本可行 □ 不可行
(4) 有何特色和创新之处	□ 具有很强的创新性 □ 具有一定的创新性 □ 创新性不明显
(5) 研究生的研究基础、实验和经费条件是否适合本选题的研究	□ 适合 □ 基本适合 □ 不适合
(6) 不足之处和需改进的方面	
(7) 其他方面	

开题报告结果（请划√）：

□ **通过**

□ **不通过**

组长签字：

年　　月　　日

六、学科审查意见

学科对开题报告的意见：

学科负责人签字：

年　　月　　日

七、学院审查意见

学院对开题报告的意见：

主管（副）院长签字：

年　　月　　日