电子科技大学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER THESIS

论文题目 健康大数据预处理技术及其应用

学 科 专 业 计算机应用技术

学 号 201421060321

作 者 姓 名 尤婷婷

指 导 教 师 卢光辉 副教授

分类号 _____ 密级 _____

UDC [注1] _____

学　位　论　文

健康大数据预处理技术及其应用

（题名和副题名）

尤婷婷

（作者姓名）

指导教师 卢光辉　副教授

电子科技大学　成　都

（姓名、职称、单位名称）

申请学位级别 硕士　学科专业 计算机应用技术

提交论文日期 _____ 论文答辩日期 _____

学位授予单位和日期 电子科技大学　2017 年 6 月

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

Research and Application of Preprocess Technology in Health Data

A Master Thesis Submitted to

University of Electronic Science and Technology of China

Major: **Computer Application Technology**

Author: **You Tingting**

Supervisor: **A.P. Lu Guanghui**

School: **School of Computer Science & Engineering**

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名：_____ 日期：　　年　　月　　日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名：_____ 导师签名：_____

日期：　　年　　月　　日

摘　要

随着信息科技的快速发展，人类社会开始步入创造和挖掘知识的信息革命时代。信息技术逐渐应用于电信、金融、教育、电子商务甚至政府决策等领域。而在国家全面建设医疗信息化的浪潮中，将大数据技术应用于与民生息息相关的医疗健康领域成为当下发展的一个热点。

由于医疗数据的特点如海量、高维度、不规范等，对医疗健康信息进行数据预处理是挖掘分析前的一个不可或缺的重要步骤。通过预处理分析不仅可以提高数据挖掘的质量，还能从一定程度上提高挖掘效率。本文结合现有技术，分别对两个医疗健康数据集进行预处理分析并对部分方法进行技术改进，主要工作如下：

（1）“人口死亡”数据集预处理方法的研究与改进。对该数据集进行特点分析并对其进行合适的预处理分析。着重研究采用随机森林算法对数据集“死亡方式”属性的缺失值进行填补。由于该数据集的非平衡性大大影响随机森林填补缺失值的效果，因此采用过采样技术 SMOTE 算法对数据集进行平衡性改善；并针对 SMOTE 算法存在的缺陷提出创新性改进。经过实验表明，使用改进后的 SMOTE 算法对数据集进行平衡性改造后，随机森林缺失值填补效果更佳。

（2）“癫痫病脑电波”数据集预处理方法的研究与改进。研究脑电波数据集预处理相关技术；并着重研究局部线性嵌入算法对脑电波频域信号进行降维。针对局部线性嵌入算法选择邻域点过大时造成的“短路边”问题，提出一种基于 K-Means 和均值的自适应选择方法。经过实验对比分析，改进后的局部线性嵌入算法具有更好的降维效果并具有良好的推广能力。

（3）对两个健康数据集预处理分析进行设计与实现。结合各自特点，将预处理技术及其相关改进应用于“人口死亡”数据集和“癫痫病脑电波”数据集分析中，为下一步的数据挖掘工作提供有效的高质量数据集。经实验表明，通过预处理后的数据集挖掘质量更佳且效率更高。

关键词：数据挖掘，医疗信息，预处理，随机森林，局部线性嵌入

ABSTRACT

With the rapid development of information technology, human society has entered the information revolution era on the basis of creation and mining. Information technology is gradually applied in telecommunications, finance, education, e-commerce, and even the government decision-making field, etc. In the tide of national construction of medical information, the Big Data technology applied to medical and health field which is closely related to the livelihood of the people becomes a hot spot in social development. Due to the characteristics of medical data, such as mass, high dimension, non-standard, preprocessing of health data is an indispensable step before data mining. Data preprocessing can not only improve the quality of the data mining, but also can improve the mining efficiency to some extent. Combining with the existing technology, firstly we analyze the key technology in data preprocessing and make some technical improvement, and then apply them into two practical medical health data sets. In this thesis, the main contents are:

1. The research and improvement of preprocessing technology in the "population death" dataset. We analyze the characteristics of dataset and then carry out the suitable preprocessing methods. In this thesis, we study random forest emphatically and then use it to fill the missing value of "death" attribute in this data set. In the process of filling the missing value by random forest, the imbalance of data set has a huge impact on the result, therefore we use one of oversampling technology--SMOTE to improve this dataset. Besides, we put forward innovative improvement aiming at the existing defects of SMOTE algorithm. Through experiment, it shows that missing value filling effect is better by using the improved SMOTE before random forest.

2. The research and improvement of preprocessing technology in the "epileptic eeg" dataset. We analyze related preprocessing methods of eeg, and emphatically study the locally linear embedding algorithm for dimension reduction of frequency domain signals. For the defect of selecting neighborhood in this algorithm, we put forward an adaptive selection method based on K-Means and mean in the step of neighborhood selection. By experimental comparison and analysis, it shows that the improved locally linear embedding algorithm is better and easy to popularize.

3. The design and implementation of data preprocessing in two health datasets.

Combined with their own characteristics, we apply proper preprocessing methods and make some improvements into "population death" dataset and "epileptic eeg" dataset for the reasonable and effective datasets in the next step of data mining. Approved by the experiment results, it shows that we have a better mining quality and efficiency after data preprocessing.

Keywords: data mining, medical information, preprocess technology, random forest, locally linear embedding

目 录

第一章 绪 论

1.1 课题研究的背景与意义

现代信息科技快速发展，人类社会开始迈入以海量数据作为基础，利用信息技术进行知识创造和信息挖掘的革命时代[1]。在信息时代的交互中，产生的海量数据反应了人们的生活习性、社会规律、自然规律等。大数据时代，数据俨然成为了一种宝贵的资产，在经济建设、社会管理、产业发展及人类生命健康等方面都起到了重要的作用[2]。而医疗领域作为一个以服务社会为目的公共事业，与人民的身体健康息息相关。因此，大数据技术在医疗领域的应用逐渐引起人民与政府的高度重视。

显然，社会的需求、时代的发展进一步促进了信息技术在医疗健康领域的发展[3]。医疗数据与信息技术的有效结合，可以在以下几方面发挥积极作用：

（1）服务居民。疾病重在预防，健康教育和指导服务系统不仅可以为居民推送健康保健指导，还可以给相关疾病例如心脑血管等提供详细的症状分析和有效预防措施。

（2）服务医生。大数据对相关疾病的挖掘分析结果，可以进一步辅助医生进行临床诊断与决策。

（3）服务科研。临床医学实验中常伴随着大量复杂的实验数据，合理地分析挖掘这些数据可以辅助科研人员识别相关疾病。

（4）服务管理机构。例如，通过对相关信息的大数据分析可以检测医疗诈骗信息、对药品进行合理定价、及时发现流行疾病并采取相关预防措施等。

显然，医疗行业是大数据技术应用的重要领域之一。通过对海量的医疗数据资源的收集及进一步的挖掘可以更好地服务居民、医院、科研和教学服务。医疗健康数据无处不在：患者基本信息、就诊信息、医疗影像、检查报告、健康档案、医保结算信息等。日积月累，这些丰富的医疗数据给后期的数据挖掘提供了真实可靠的信息。然而，现实生活中的数据并不总是那么完美。在医疗信息收集中，由于数据库设计差异、填写不规范、内容错填漏填、数据管理维护不当等因素造成数据质量低下。这样的数据将导致挖掘结果低质量或严重偏颇。同时，维度高达成千上万的数据不仅具有复杂的几何结构、难以检测，还给人们的数据处理工作带来了相当大的困难。因此，预处理工作在医疗健康数据挖掘分析中具有举足轻重的地位。

在一个完整数据挖掘项目中，真正的挖掘工作仅占挖掘分析总工作量的10%左右；而数据预处理分析则通常要花费70%的时间[4]。数据预处理不仅可以提高数据

挖掘质量还能降低挖掘复杂度提高效率，其重要性不言而喻。

1.2 国内外研究现状

1.2.1 医疗大数据的国内外研究现状

20 世纪中叶，信息技术便开始应用于国外的医疗系统中。在半个多世纪的信息技术发展与社会需求的推动下，数据挖掘技术渐渐地在临床、影像以及医院的决策管理等方面取得一定的成效。如使用图像检索等技术辅助医生诊断相关皮肤疾病、肿瘤等。2009 年，互联网巨头谷歌也通过人们在网上最频繁检索的词条记录成功地比美国疾病控制与预防中心提前一到两周预测到了甲型 H1N1 流感的爆发，至此数据挖掘在医疗领域也渐渐崭露头角。

在国际医学研究发展上，美国、欧洲、日本已经走在了世界的前列。而随着信息技术全球化和转型医学的兴起，我国的数据挖掘在医学研究领域也实现了历史性的跨越。如 39 健康网，患者可以在网上提供病症数据从而获得初步诊断。同时，结合智能穿戴设备实时获取用户的健康信息，可对用户的健康做出预测和提醒服务，例如提醒用户注意饮食和规律生活作息。除此之外，首都医科大学附属北京安贞医院正尝试采用大数据技术研究心血管疾病与环境之间的关系。在国家“863”计划 2015 年项目申报指南中，医学大数据标准化和集成、医学大数据表述搜索与存储访问、融合技术等服务技术都涉及其中，旨在构建大数据中心和知识库[5]。在人民迫切需要与政府的助力推动下，数据挖掘在医疗领域中的应用研究正在如火如荼地进行中。

1.2.2 医疗大数据预处理技术的国内外研究现状

现实世界中，数据常因各方面侵扰而使得数据挖掘的质量降低。同时，在数据挖掘中，如果数据的维度特别的高，这也会大大地增加数据挖掘的难度和时间，降低数据挖掘的效率。当然在医疗领域，数据依然存在着以上的问题。而通过数据预处理可以有效提高数据质量，节约大量的时间和空间，最终提高挖掘的效率和质量。因此，数据预处理技术同样也在医疗领域得到了相当的重视。

目前，有的医院对病人电子病历 EMR 数据进行数据集成和清理并提取信息，从而有效提高对直肠癌预测的准确率；同时，也有利于直肠癌的早期发现和干预治疗实践[6]。而另一个案例，则是收集医院重症监护病房的数据预测脓毒性休克患者的存活与死亡结果。由于该类数据往往是不定期记录的，因此存在缺失值过多和采样时间不均匀的问题。于是采用数据预处理技术填补缺失值并且解决采样时间不均匀的问题，提高了挖掘数据的质量[7]。再如对于病人的体检报告，医生结合

检查者的个人信息及以往身体检查记录，可以分析出检查者的身体状况并为其提供相关的预防措施和健康指导。然而，信息技术与医疗领域之间的学科差异导致现有体检数据库存在着许多问题，如异常值较多、基本信息缺失、体检项目名称不统一、指标参考值度量单位不同、唯一标识码缺失等等，这一系列问题都会影响数据挖掘的质量。因此，在使用这些历史体检数据进行挖掘前，对数据进行预处理是至关重要的[8]。在医学图像方面，自动诊断皮肤癌是最具挑战性的问题之一。它可以帮助医生决定皮肤黑素瘤是良性的还是恶性的。通过数据预处理去除图像无关的噪音和不必要的背景图像，从而光滑图像，有效提高图像的质量[9]。

然而就国内现状而言，数据挖掘在医学研究领域的发展仍然充满机遇与挑战。首先，由于各个医院之间都是独立的，从而对患者的数据共享产生了阻碍，使得转换利用率低下；因此需要加强医学研究大数据的整合与共享[5]。其次，由于信息技术与医学研究两门不同学科间的鸿沟，针对两者融合的研究较少；因此应该加强信息技术与医学研究的融合。最后，国家政府也应该积极引导设立专项资金向医学研究大数据项目倾斜，加强专业人才培养。

总之，通过对医疗健康数据的挖掘，不仅可以辅助诊断疾病、检测潜在疾病、辅助医生诊断；还能服务民生、医学单位和科研机关。在医疗数据挖掘中，数据预处理成为影响数据挖掘任务成功与否的关键因素之一[10]。因此医疗健康领域的预处理技术，是一个亟待深入研究的领域。

1.3 论文的研究内容与创新

本论文主要对医疗相关的健康数据挖掘的预处理核心技术进行研究，并将其真正运用到实际的健康数据挖掘分析中。实际的健康数据内容包括从国家卫生局收录的人口死亡数据以及从医院获取的癫痫病患者的 EEG 脑电波数据。将数据挖掘技术应用于与民生息息相关的医疗领域，是迈向国家医疗信息化的一个重要步骤。

1.3.1 本文的研究内容

随着近几年大数据的快速发展，将数据挖掘技术应用到医疗行业中是国家医疗高度信息化建设的必然趋势。而在健康数据的挖掘中，对数据进行合适地预处理分析是尤为重要的。因此本论文的工作主要从以下几个方面进行展开：

1、数据挖掘中数据预处理的关键技术的研究。本文首先分析了数据预处理的重要性，然后对预处理工作中的四个核心步骤即数据清理、数据集成、数据归约和数据变换中涉及的相关处理技术进行研究，为之后医疗健康数据的预处理分析

实际应用做系统的、完善的理论知识储备。

2、人口死亡数据集的预处理分析技术研究与应用。本论文的第一个实际应用是从国家卫生局收录的人口死亡数据集的预处理分析；将数据预处理分析的理论技术应用于实践中，并重点研究使用随机森林算法对数据集缺失值进行填补。又根据人口死亡数据集的不平衡性特点，对优化非平衡数据集的 SMOTE 算法进行改进，从而达到随机森林填补数据集缺失值的效果更佳的目的。

3、癫痫病患者 EEG 脑电波的预处理分析技术研究与应用。本论文的第二个实际应用是癫痫病患者 EEG 脑电波数据集的预处理分析；通过各种方法的实验结果对比选出最有效的预处理方法，并重点研究使用局部线性嵌入算法对脑电波频域信号进行降维。针对局部线性嵌入算法中存在的缺陷与不足对算法进行改进并应用于脑电波数据集分析中。

1.3.2 本文的创新点

本文的主要创新点在于：

1、使用随机森林填补非平衡数据集缺失值的改进。使用随机森林对“人口死亡”数据集中的死亡方式属性进行填补。由于该数据集的非平衡性严重影响填补结果，因此采用过采样技术 SMOTE 算法对数据集进行非平衡性改善。传统的 SMOTE 算法受噪声数据和边缘数据影响较大，因此在合成新样本时，本文对 SMOTE 算法做出了创新性的改进，使得合成的新样本与原始数据集更加切合，更好的改善了非平衡数据集，为随机森林的缺失值填补提供高质量数据。

2、使用局部线性嵌入算法对脑电波频域信号进行降维的改进。在对癫痫病患者脑电波频域信号进行预处理时，采用了局部线性嵌入算法对其进行有效地降维。但局部线性嵌入算法的邻域点数选择不合适会导致低维嵌入时效果较差。因此，针对这一问题，本文提出了一种与聚类方法相结合的邻域点选择方法，即基于 K-Means 和均值限制的邻域点选择方法。使用改进后的方法可以自适应地选择邻域点，有效提高了对脑电波频域信号降维的效果。

3、两个健康数据集预处理分析的设计与实现。分别对两个数据集的特点进行分析，从而设计合适的数据预处理步骤，并将其真正应用于两个数据集即“人口死亡”数据集和“癫痫病脑电波”数据集中。经实验结果分析表明，对数据集进行数据预处理后，其挖掘效率和挖掘质量得到显著提高。

1.4 本论文的结构安排

本文首先对数据挖掘预处理技术的理论知识进行研究，然后根据两个实际数据

集（即人口死亡数据集和癫痫病脑电波数据集）的特点对其采取合适的数据预处理，并针对其中的某些关键技术进行探讨分析并提出创新性的改进方法。本论文总共分为六个章节，具体的结构安排如下：

第一章，首先介绍本课题的研究背景与意义，然后对数据预处理技术的国内外研究现状进行简要介绍，最后对本文研究内容和创新点进行概述。

第二章，介绍数据预处理的常用关键技术。对数据清理、数据集成、数据归约、数据变换的常用关键技术进行了简要介绍。

第三章，介绍随机森林相关技术用以填补“人口死亡”数据集的缺失值。研究随机森林相关技术并着重对处理非平衡数据集时的方法进行研究与改进。

第四章，介绍局部线性嵌入算法用以对“癫痫病脑电波”数据集进行降维处理。研究对比各降维技术，并着重对局部线性嵌入算法进行优化改进。

第五章，分别对“人口死亡”数据集和“癫痫病脑电波”数据集的数据预处理进行设计与实现。对数据集进行需求和特点分析，设计框架结构；并对数据预处理核心模块的实现进行详细阐述；最后，进行结果对比展示与分析。

第六章，对本文实现的预处理技术进行总结。回顾数据预处理工作，并对所做工作的优缺点进行分析，最后对医疗健康领域的数据挖掘的未来发展做出展望。

第二章 数据预处理关键技术

2.1 医疗数据预处理的必要性

数据挖掘技术以一种全新的概念改变着医疗领域。随着医疗行业的转型和大数据的发展，医院可收集大量的医疗数据，如何从多样化的海量数据中挖掘有效信息是医疗健康领域的一个重要研究方向。但是，医疗数据的挖掘与其他领域又有所不同，一定程度上是由数据特性的不同导致的。据医疗历史数据和相关资料的分析，医疗数据主要具备以下特点：

（1）海量性

医院信息系统的建立和不断完善，使得医院能够更加方便地收集和存储医疗数据。据统计，一家市级医院每年的门诊量就高达 400 万人次，由此可想医院可收集的医疗数据之多。除此之外，存储的部分文件也是巨大的。例如仅一个 CT 图像存储大小约为 150MB，而一个标准的病理图的存储空间高达 5GB。

（2）异构性

医疗数据包含了多种类型的数据。如体验、化验结果为纯数据类型，B 超、X 线为图像类型，脑电波、心电图为信号类型；检测报告、过敏史为文字类型；除此之外还包括用于科普教育或咨询的动画语音视频信息等。这些具有专业性的信息都成为了区别于其他领域的最显著特征。

（3）不完整性

由于医疗信息的收集和处理阶段的相互脱节，使得医疗数据库并不能对所有的疾病信息做到全面反映。又由于医疗数据大量来源于人工记录，因此数据很可能会出现偏差和残缺问题；加之数据的不清晰表达、记录本身的不确定性（病例和病案尤为突出）也造成了医疗信息的不完整性。

（4）冗余性

医院每天收集的大量数据可能会包含重复、无关紧要的数据；甚至有的数据记录互相矛盾。

（5）不规范性

目前，由于医疗领域系统的复杂性和庞大性，许多名词和概念都还无法做到统一标准规范化。此外，在诊断过程中，医生的知识背景、诊断方式、经验和病情描述的差异也会使得医疗信息常呈现出一些主观性。因此医疗数据通常是不规范的。

基于以上医疗健康数据的特点表明：收集到的医疗数据往往是低质量的，非

常不利于进行数据挖掘。因此，在数据挖掘之前对医疗数据进行预处理是非常有必要的，它是直接影响到挖掘分析工作成功与否的关键因素之一。目前，数据预处理工作主要包括数据清理、数据集成、数据归约和数据转换。

2.2 数据清理

现实世界的数据常呈现出信息不完整、表达信息不一致和受噪声影响大等特点。数据清理则通过识别或删除离群点来光滑噪声数据，并填补缺失值从而“清理”数据。因此，数据清理的主要任务为：缺失值处理、光滑噪声数据[4]。

2.2.1 缺失值处理

数据缺失是一个存在于许多领域且无法避免的复杂问题。在数据挖掘过程中，空值的存在会引起很多问题。例如，空值的存在会导致系统丢失大量有价值信息；其次，系统中蕴藏的确定性信息更难以把握，不确定性更加显著；再次，包含空值的数据降低挖掘质量，导致不可靠的输出。因此，采用合适的方法对缺失值进行填补是必要的。在填补缺失值的处理上有以下几类处理方法。

（1）删除元组。删除元组是一个最简单直接的方法，该方法通过删除存在遗漏信息的对象并整合剩余对象，从而得到一个完整的信息表。当数据信息表中含有缺失值的对象比例很小时，常使用该方法处理。然而，这种方法虽然保持了数据的完整性，却在减少历史数据时丢弃了隐藏在数据对象中的信息，造成了资源的浪费。除此之外，当遗漏数据在整个数据集中占据比例较大，特别当这些数据呈现非随机分布时，使用该方法可能导致数据的偏离，最终导致挖掘质量低下。

（2）人工填写缺失值。顾名思义，因为用户本人最了解自己的信息，因此使用该方法填充的数据真实可靠，具有数据偏离最小，填充质量最优的优点。然而，在医疗信息库中，病人的临床检验结果并非都能在特定时间内轻易得到，因此该方法实现较为困难。同时，当待填充数据规模大、空值较多时，该方法耗时较长。

（3）中心度量填充。该方法使用现存数据中的多数信息来填补缺失值。信息表中的属性按类别可划分为非数值属性和数值属性。当空值为数值型时，则根据该属性取值的平均值来填充缺失值；当空值是非数值型时，则根据统计学原理，使用该属性取值频次最高的数据来填补缺失值。

（4）多重填补[11]。多重填补是以贝叶斯估计为基础的，它的主要思想为：待填补值是随机分布的，并且这些信息可以从已观测到的数据得到。具体步骤为：首先，为每个空值产生一套可能的填补值，分别使用这些值进行缺失值填补，从而产生若干个完整数据集；然后，使用挖掘技术对每个填补后的完整数据集进行

挖掘分析；最后，通过分析各填补数据集的结果，选出最佳填补方式。

（5）使用最可能的值填充。随着人们对缺失值处理方法的研究的深入，逐渐将数据挖掘的方法应用于填补缺失值上。例如用回归、贝叶斯形式化方法，或者决策树、随机森林确定最可能的缺失值。其基本思想是通过建立应变量 Y 和自变量 X 的模型来预测确实变量 Y 中的缺失数据。通过这类方法得到的估计值往往更加接近真实值，但构造和评估模型的过程比较复杂，需要对模型进行评价。本文的下一章将会对这类相关技术展开详细阐述。

2.2.2 光滑噪声数据

同时，由于数据收集的填写不规范、数据管理维护不当等因素往往会使数据库受噪声的侵扰，这些噪声数据直接影响着挖掘分析结果。目前，主要存在以下的数据光滑技术：

（1）分箱。分箱方法的主要思想为：每一个数据与它的“近邻”数据应该是相似的，因此将数据用其近邻（“箱”或“桶”）替代表示既可以光滑有序数据值，还能在一定程度上保持数据的独有特点。图 2-1 展示了一个分箱实例。在该例中，先对年龄数据 age 按升序排列，并将其均匀划分到等频箱中。当使用箱均值光滑时，用每箱计算的均值替代箱中每一个值。例如，箱 1 中的均值为 9，则该箱中的每个值都被 9 替换。而使用箱边界光滑时，则使用箱中的最大值和最小值作为箱边界，然后用其替换箱中的数值。

划分为（等频的）箱：	用箱均值光滑：	用箱边界光滑：
箱1： 4，8，15	箱1： 9，9，9	箱1： 4，4，15
箱2： 21，21，24	箱2： 22，22，22	箱2： 21，21，24
箱3： 25，28，34	箱3： 29，29，29	箱3： 25，25，25

图 2-1 数据光滑的分箱方法

（2）回归。回归技术是通过一个映像或函数拟合多个属性数据从而达到光滑数据的效果。线性回归则是寻找一条“最佳”直线来拟合多个属性，从而实现使用其中的某些属性预测其他属性。

（3）离群点分析。聚类可以将相似的值归为同一“簇”中，因此主要使用聚类等技术来检测离群点。如图 2-2 所示，显示的是顾客在城市中的位置，通过聚类可以检测到在簇集合之外的点。

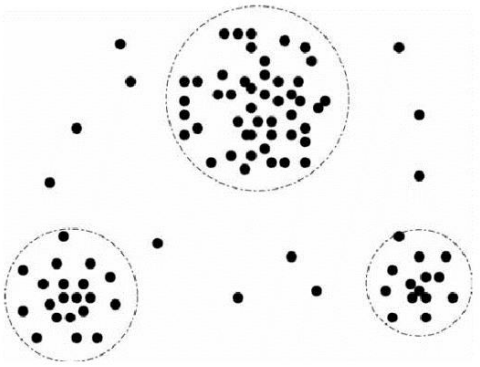


图 2-2 顾客在城市中的位置，显示了 3 个簇，离群点为簇之外的值

总而言之，数据清理是一项比较繁重的任务，它会随着数据自身的特点和挖掘需求采取相应的措施。因此，并没有一个步骤一致的数据清理过程。

2.3 数据集成

数据挖掘常需进行数据集成——将来自多个数据库存储的数据统一存放到一个统一的数据存储中。现今各个医院设计的系统并非统一管理，因此每个医院对自己的数据库有着独特的管理方式，这便使得合并数据库实现信息共享变得十分困难。很明显，直接将两个不同设计不同的数据库合并到一起是不可行的，这可能会造成数据集的冗余和不一致，因此如何匹配多个数据源的模式和对象，是数据集成解决的主要问题。在实际应用中，数据集成解决三类问题：实体识别、冗余和相关分析以及数值冲突的检测与处理。

2.3.1 实体识别问题

数据集成涉及许多问题，实体识别则是寻找匹配来自多个数据信息库的等价实体。例如，一个数据库中的属性名 patient_id 与另一个数据库中的属性名 patient_number 表示含义是否相同。每个属性的元数据包括属性名、现实含义、数据类型、取值范围，以及处理零或空白时的空值规则。元数据的统一设计不仅可以有效避免模式集成的错误，还能在变换数据时也起到一定作用。例如，对于 sex 这个属性，在一个数据库中用 male 和 female 表示，而另一个数据库则用数字 0 和 1 来表示。在对多个数据库进行集成时需注意的是：相同的属性名并不意味着相同的数据结构或含义。例如 discount 这个属性除了表示折扣率以外，还可用于表示商品是否处于打折状态；如果在集成之前这些差异未被发现，则会为之后的数据挖掘造成困难。

2.3.2 冗余和相关分析

当对多个数据库集成时，常会出现数据冗余现象。例如一个人的出生年份可以通过年龄导出，那么出生年份这个属性就是冗余的。或者对于同一现实实体，不同数据库有其相对应的属性，因此集成也会造成数据冗余。分析冗余有很多方法。首先，可以将数据进行可视化处理，将数据点绘制成图表后趋势和关联会变得清晰起来。除此之外，冗余还可以通过相关性分析方法检测。对于标称数据，可以使用卡方检验；对于数值属性，可以使用以下两种方法进行分析。

（1）协方差（Covariance）。协方差在统计分析或数据挖掘中常用于衡量两个变量的总体误差，计算公式如下：

$$Cov(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1} \tag{2-1}$$

且根据协方差的不同数值具有以下结论：

$$Cov(X,Y)=\begin{cases}>0 & X,Y\text{呈正相关}\\ <0 & X,Y\text{呈负相关}\\ =0 & X,Y\text{相互独立}\end{cases} \tag{2-2}$$

（2）相关系数 （Correlation）。当我们面对多个变量时，无法通过协方差来说明哪两组数据的相关性最高。因此需使用相关系数来衡量和对比相关性的密切程度。其计算公式如下：

$$r_{xy}=\frac{s_{xy}}{s_x s_y} \tag{2-3}$$

其中 r_{xy} 表示样本相关系数；s_x，s_y 分别表示 x 和 y 的样本标准差；s_{xy} 表示两样本的协方差。相关系数的优点是可以通过数字对变量的关系进行度量并带有方向性，它的取值为 $[-1,1]$，数值表示的含义与协方差类似。

综上，通过相关性分析可删除冗余数据，以达到精简数据集，降低计算复杂度的目的。

2.3.3 数值冲突的检测与处理

对于现实世界的同一实体，由于表达方式、尺度标准或编码的不同常导致元数据的巨大差异。例如，对于身高这一属性，在一个系统中可能以“米”作为度量单位，而在另一个系统中则可能以“厘米”作为度量单位。又如在大学的课程评分系统中，有的学校采用 A+~F 对成绩评分；而有的则采用数值 1~10 评分。于是在对这两所学校进行数据库合并时，应该将两个系统的评分制度做统一处理，以便进行进一步的数据挖掘。

数据语义的多样性和每个属性的元数据对数据集成提出了巨大的挑战。但是将多个数据源的数据进行集成可以提供更多的参考数据，有利于挖掘出更有价值的信息。因此，谨慎地进行数据集成可以降低数据集的不一致性和冗余，有助于提高之后挖掘过程中的精度和速度。

2.4 数据归约

在当今信息时代，数据量的迅猛扩增是挖掘技术发展的驱动因素之一。然而，对海量数据的分析和挖掘不仅增加了技术的复杂度也大大延长了挖掘的时间，这是数据挖掘研究面临的一个巨大挑战。针对这一类问题，数据归约技术应运而生。该技术使用精简的数据集来代替原始的庞大数据集；它虽比原始数据集小得多，但却良好的保持了原始数据的完整性与独有特性。对于复杂庞大的数据集，数据归约步骤必不可少。它不仅可以有效降低挖掘复杂度，减少挖掘时间，还具有良好可靠的挖掘质量。

2.4.1 维归约

长久以来，数据库中的属性不断地增加，这为数据挖掘提供了更加丰富、细致的信息。显然随着数据信息量及维度不断地增大，造成的“维度灾难”成为了数据挖掘工作者们面临的一大难题。维归约技术主要解决数据维度过于庞大的问题，它的主要思想是减少随机变量的个数。根据是否数据变换，将其分为特征提取和特征选择[12]。

特征选择是一种对属性子集选择的方式，其目的是检测对挖掘结果无影响或影响较小的属性并将其删除。它的好处包括：便于理解和可视化数据，降低数据集维度，从而降低计算及存储压力。现在有三种主流方法：过滤式，包裹式，嵌入式。

（1）过滤式。这种方法的关键就是找到一种能度量特征重要性的方法，比如上节中提到的卡方检验，相关系数等。该方法的主要缺点是忽略了特征之间可能存在的相互依赖关系，也忽略了对冗余特征存在和不明显特征的的考虑。

（2）包裹式。这类方法的核心思想是：给定某种模型及预测效果评价的方法，然后针对特征空间中的不同子集，计算每个子集的预测效果，选择效果最好的特征子集作为最终特征子集。不过由于包裹式方法要求针对每一个特征子集重新训练模型，因此计算量还是较大的。

（3）嵌入式。该方法是在模型的训练过程中对特征进行选择，比如决策树在分枝的过程中，就是使用的嵌入式特征选择方法，其内在还是根据某个度量指标

（如信息熵）对特征进行排序。

特征提取则是寻找一个映像或函数将原始数据集中的高维数据转换成低维数据[13]，即通过更少的维度来表示数据内部的本质结构特征。常用的降维方法有小波变换和一系列的降维算法。

小波变换[14]。小波变换使用不同频率的小波函数的和拟合原始信号，通过保留大于某个设定阈值的小波系数，从而达到降维的目的。小波变换技术广泛应用于图像处理、时间序列数据分析、计算机视觉等。

为了有效地处理复杂的高维数据，解决“维度灾难”问题，数据降维技术研究得到科学界的高度重视。依据数据间的关系，可将其分为线性降维和非线性降维。

线性降维技术通常假定数据的各变量是相互独立的，通过线性降维把数据投影到低维线性子空间。主要的线性降维方法有：主成分分析（PCA）[15]，独立成分分析（ICA），线性判别分析（LDA）。

非线性降维比线性降维稍复杂些，它假定各个属性间呈强相关性，并具有高度的非线性特点；例如文本数据、音频数据、视频数据、图像数据等。这些数据结构复杂，因此需要采用非线性降维。比较流行的非线性降维方法有：局部线性嵌入（LLE），等距映射算法（ISOMAP），拉普拉斯特征映射算法（LE）等。本文的第四章将对降维算法进行展开研究。

2.4.2 数值归约

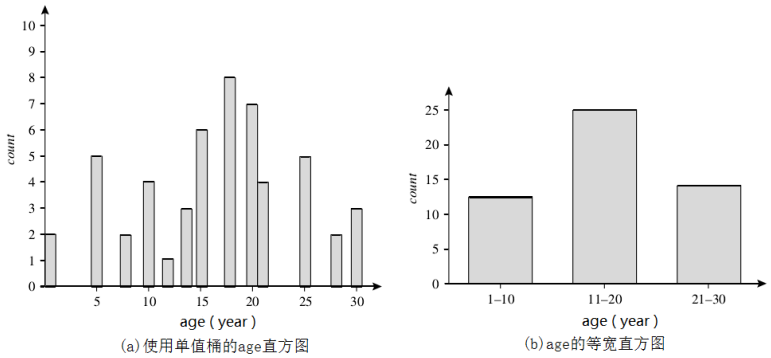
数值归约的主要思想是用较小的数据来替代原数据从而达到减小数据量的目的。相关技术可分为参数方法的和非参数方法。在参数方法中，较常用的是回归和对数-线性模型；非参数方法包含聚类、直方图等。

在参数方法中，主要是对应变量进行建模，从而找到一个模型可以将多个变量映射成一个变量，达到减小维度的目的。回归和对数-线性模型的应用具有局限性，常用于稀疏矩阵。对于高维数据，回归比较适合计算密集的；而对数-线性模型有比较好的可伸缩性。

聚类技术则是按照某种规则将对象划分为群或簇，使得簇中对象彼此“相似”，且与其他簇中对象“相异”[16]。在数值归约中，通常将 n 维数据看做 n 维空间的点，再通过聚类将这些数据划分到不同的簇中，最终使用簇标号来代替这 n 维元组，从而达到缩减数据集的目的。

直方图使用分箱来模拟近似数据分布。首先，将连续的属性分割成不相交的区间，这里称作桶，即每个桶代表了一个属性值的连续区间；然后将属性对应的

数值分别划分到这些单桶中。如下图 2-3 所示：



(a) 使用单值桶的age直方图　　(b) age的等宽直方图

图 2-3 age 的分箱操作直方图

图(a)为使用单值桶的 age 的直方图，每个桶代表了一个 age 值。在对数值进行规约时，分别使用一个桶代表 age 属性的一个连续年龄值域，即如图(b)显示的。每个桶代表年龄的一个连续 10 岁区间。这样将繁琐连续的属性通过区间划分为了更小数据集合，使得数据更加简洁。

2.5 数据变换

每个系统都有自己独有的数据库管理方式，需求的不同会导致数据库设计和管理的差异，即元数据的数据类型和允许取值范围会根据实际情况而调整。因此在数据集成时也应该对数据属性进行统一的变换。除此之外，数据挖掘技术有时也会对数据格式有特定限制。比如在用决策树进行分类时，就必须保证属性都是数值型的。因此，在进行数据挖掘前，需对数据集进行相应的数据变换。常用的的变换策略介绍如下。

2.5.1 数据规范化

数据规范化的目的是将数据按比例缩放，使得属性之间的权值适合数据挖掘。例如，统计身高信息的度量单位是有所不同的，若在数据挖掘中把 height 属性的度量单位从米变成英寸，则可能导致完全不同的结果。一般而言，度量单位的不同将导致属性的值域取值范围不同，也因此该属性的权重在分析过程中会发生相应的变化。在对指标参与评价的计算中，需要对指标的度量单位和取值范围进行规范化处理，通过合适的函数变换将其映射到设定的数值区间。常见的数据规范化方法包括以下三种。

（1）最小-最大规范化。该方法对数据的变换呈线性。若 max_A 和 $\min_A$ 分别表示属性 A 的最大值和最小值。则使用最小-最大规范化计算公式：

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \tag{2-4}$$

可将属性 A 的值域转换到区间[new_min$_A$，new_max$_A$]。采用这种线性变化可保持数据与原始数据之间的联系。但当输入的实例落在 A 的原始数据值域之外时，通过最小-最大规范化处理后的值将可能有“越界”危险。

（2）z-score 分数规范化。该方法可以将较大的值转换成较小的值。计算公式如下所示：

$$v' = \frac{v - \overline{A}}{\sigma_A} \tag{2-5}$$

公式中，$\overline{A}$ 为属性 A 的均值而 σ_A 为属性 A 的标准差。例如，消费额 money 的均值和标准差分别为 28000 人民币和 12000 人民币，则使用 z-score 分数规范化计算公式 $\frac{36000\text{-}28000}{12000} = 0.667$ 后，money 属性值从 36000 转换成 0.667。因此，我们将很大的数值通过 z-score 转变成较小的数据，在一定程度上减小了之后数据挖掘的复杂度。

（3）小数定标规范化。该方法是通过移动数值的小数点位置来达到缩放效果的规范化处理，通过小数定标规范化后的值域转变为[-1，1]。计算公式如下：

$$v' = \frac{v}{10^j} \tag{2-6}$$

通过公式可以看出，若要将数值转换到[-1，1]这个区间，则小数点的移动位数由属性的最大绝对值决定，因此 j 的取值为使得 $\max(|v|) < 1$ 的最小整数。

数据规范化中的 z-score 分数规范化方法和小数定标规范化方法将对原始数据产生巨大改变。因此，考虑到新数据的一致性处理，需注意在转换过程中对规范化参数进行保留。

2.5.2 数据离散化

数据离散化是将数值属性的原始值用区间标签或者概念标签替换的过程[17]，它可以将连续属性值离散化。以学生的考试成绩为例，如图 2-4 所示：

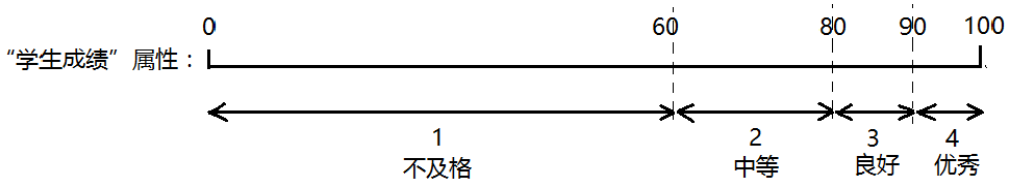


图 2-4 学生成绩的离散化处理

从图中可以看出，“学生成绩”属性是一个分布在 0 到 100 的连续数值。在图中，将其属性分割成 4 个区间：0 到 60 分为‘不及格’，用数值 1 来表示；60 到 80 分为‘中等’，用数值 2 表示；80 到 90 分为成绩‘良好’，用数值 3 来表示；而 90-100 分归为‘优秀’，用数值 4 来表示。通过这种方式，将“学生成绩”这个连续的属性值用 4 个离散的数值表示。总而言之，连续属性离散化的实质是将连续属性值转换成少数有限的区间，从而有效提高挖掘工作的计算效率[18]。

2.5.3 概念分层

概念分层的主要思想是将低层概念的集合映射到高层概念的集合[19]，它广泛应用于标称数据的转换。通常情况下，分类属性的概念分层往往涉及一组属性。可以通过专家或用户对属性进行偏序或全序的设定，从而对属性进行概念分层。如现有某个数据库需对关于地理位置 location 的属性集进行概念分层，其中属性内容包括：街道 street，国家 country，城市 city 和省份 province_or_state。首先，对每个属性不同值的个数进行统计分析，并将其按照升序进行排列；其次，根据排列好的属性次序，自顶向下进行分层。其结果如图 2-5 所示：

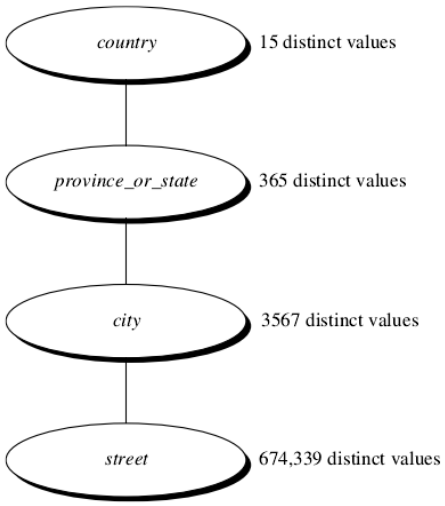


图 2-5 地理位置的概念分层

如大家的常规认识，对属性的全序排列结果为街道 street < 城市 city < 省份 province_or_state < 国家 country，即街道 street 属性在最顶层，国家 country 属性在最底层。最后，用户根据产生的分层，选择合适的属性代替该属性集。

使用概念分层变换数据使得较高层的知识模式特点突出，容易被发现。它允许在多个抽象层进行挖掘，这是许多数据挖掘应用的共同需要。

2.6 本章小结

本章首先对医疗健康数据的特点进行分析，阐述了挖掘之前对医疗数据进行预处理的重要性。之后，介绍了数据预处理的常见关键技术，主要包括：数据清理，数据归约，数据集成，数据变换。同时，针对各个关键技术的主要方法进行了简要阐述，为之后具体的医疗数据预处理方法研究做基础理论知识储备。

第三章 人口死亡方式缺失值填补技术的研究与改进

3.1 现代填补类别缺失值的技术研究

3.1.1 填补缺失值的必要性

数据缺失是一种普遍存在的现象，在系统录入过程、实验研究或是日常生活中的抽样调查，都会因为种种原因而导致数据缺失。在对本章的“人口死亡”数据集进行数据挖掘时，“死亡方式”这一属性存在着缺失值。而这些缺失的数据会在一定程度上影响数据挖掘的质量。首先，数据的缺失会导致获取信息量相应减少。其次，在对数据集进行初步分析时发现，“死亡方式”属性是一个较为重要的特征属性，对挖掘结果的影响较大。总之，“死亡方式”数据的缺失在很大程度上影响了数据的整体质量，使得统计数据的说服力降低，最终影响数据挖掘的结果[20]。

也正是这样，人们开始对缺失数据处理方法展开了深入的研究。在上一章的数据清理小节中，讲解了缺失值填补的一些技术；包括传统缺失值填补的处理方法和现代缺失值填补的填补方法。本章主要研究的是填补类别缺失值的现代处理方法。类别缺失数据就是在数据集中缺少了某些非连续值表征的类别数据。例如在性别这一栏中，男或女则是类别的数据；在卫生局健康登记表中，自杀或非自杀也是类别的数据。填补类别缺失数据即将这些缺失的类别信息通过某种方式填补从而有利于进行下一步的数据挖掘分析。简而言之，现代的填补类别缺失值的处理方法就是对数据分类的研究，即利用分类器建立分类模型从而预测并填补缺失值。

3.1.2 现代分类方法的研究

现代的分类方法主要分为基本方法和高级方法。分类的基本方法诞生于数据挖掘的早期，它主要结合了一些统计学的方法进行研究，思想也较为简单。典型主要方法有：朴素贝叶斯分类、最近邻规则分类及决策树分类。

（1）最近邻规则分类（ K-Nearest Neighbor，KNN ）。它的主要思路[21]是：在特征空间中，一个样本应该与其最近邻的样本最为相似；如果在它 k 个近邻中的大多数样本属于某一个类别，则该样本也应该同属于这个类别，并具有一定的相似性。如果样本选择的近邻点都正确分类，则使用 KNN 算法有较好的分类效果。在图 3-1 中：

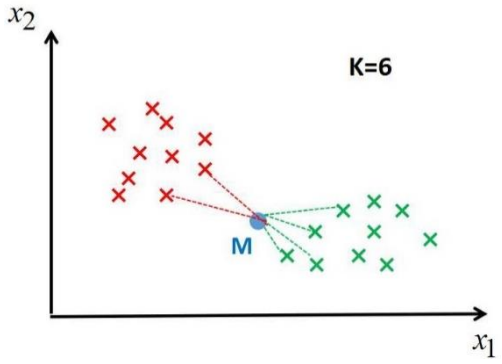


图 3-1 KNN 算法的分类实例图解

图中，当 K 为 6 时，有 4 个近邻属于绿色类，只有 2 个近邻属于红色类，则判定 M 是属于绿色类的。由于 KNN 方法不使用判别类域特性的方法而主要依赖待分类样本有限的邻近样本来确定其所属类别；因此对于类域边界重叠模糊的待分样本集来说，KNN 方法分类效果更佳。

（2）朴素贝叶斯分类（Naive Bayes，NB）。朴素贝叶斯是基于概率统计知识而形成的分类算法，它以贝叶斯定理作为理论基础。该方法主要思想是：对于给出的待分类样本，计算此项出现的条件下各类别出现的概率大小，选择最大概率所属的类别作为最后待分类项最终预测的类别[22]。理论上，朴素贝叶斯分类算法可以和高级分类方法如神经网络媲美；但在实际应用中，由于它要假设数据集中各个属性对于给定类是相互独立的，因此应用场合受到一定的限制。

（3）决策树分类（Decision Tree，DT）[23]。决策树是一种归纳学习算法。它通过一种合适的分类规则将一组无规则次序的元组用决策树的形式表示。直观上，决策树分类器就像判断模块和终止模块组成的流程图；判断模块表示对一个特征取值的判断即树的分支，终止模块表示分类结果即树的叶子。决策树使用属性选择度量算法来选择元组，并将其最好地划分成不同的类的属性；常用的属性选择度量算法有 ID3 和 C4.5。

决策树在数据集受到较大噪声干扰时，容易出现过拟合的现象；即构建好的决策树模型对训练数据的分类效果较好，但是将其运用到测试数据上分类性能则会大大降低。

高级分类技术则是较近期研究出的新的分类解决方案。它们同基本方法相比，虽然复杂度较高，却常常伴有更高的分类准确度，因此也受到科学界高度的重视和广泛的优化研究。主要的技术包括：支持向量机、人工神经网络、集成分类。

（1）支持向量机（Support Vector Machines，SVM ）。SVM 是一种以统计分析理论为基础的新机器学习方法[24]。它的基本思想是：寻找一个合理的超平面将样本数据集线性划分，当不同类样本距离超平面最小距离之和达到最大化时，则

称这个超平面为最佳超平面，而这些距离最佳超平面最近的样本则称为支持向量。如下图 3-2 所示：

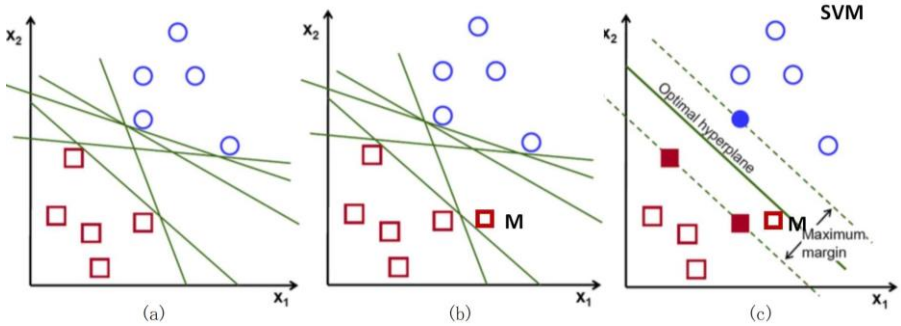


图 3-2 SVM 算法的分类实例图解

在(a)图中，多条绿色直线将数据样本的红类和蓝类进行正确地线性划分。但在(b)图出现新样本点 M 时，之前正确划分样本数据的绿色直线却根据自己的划分平面将 M 进行了错误的分类。在(c)图中，通过 SVM 划分的超平面由于保持了红类和蓝类间最近距离的最大化，所以靠近红类的新样本 M 被正确划分。

SVM 在学习能力和模型的复杂度之间寻求最佳折衷，具有良好的泛化推广能力。但因为 SVM 是通过求解二次规划问题来寻求支持向量的，因此当数据量巨大或维度较高时，支持向量机分类方法在存储和计算时对机器内存要求较高，且运算时间长。再之，目前较为成熟的支持向量机算法仅针对二分类问题[25]，在解决多分类问题的实际挖掘应用场合受到一定的限制。

（2）人工神经网络（Artificial Neural Network，ANN）。ANN 结合网络拓扑结构知识和人脑神经系统对复杂信息的处理机制，形成了一种具有预测能力的数学模型。本质上，它是一个通过大量简单元件交错连接而形成的复杂网络拓扑结构[26]。也因此神经网络呈现出高度复杂的非线性结构，在处理分类问题时准确率较高。根据拓扑结构的不同可将其分为前向网络和反馈网络。如图 3-3 所示：

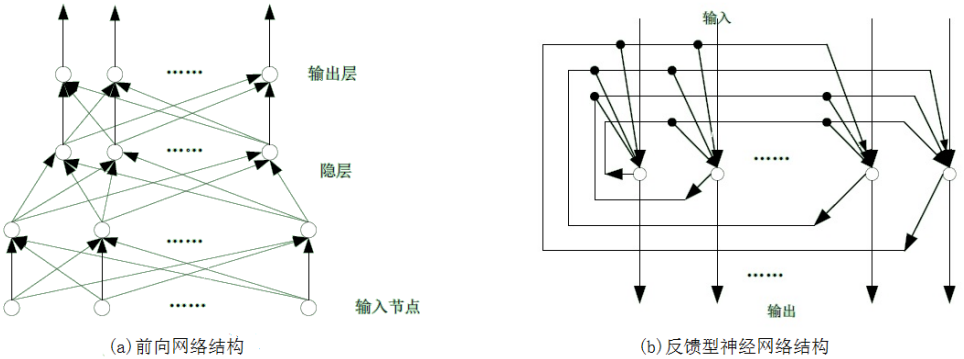


图 3-3 人工神经网络的两种拓扑结构图

在(a)前向网络结构图中网络被分为若干“层”，上一层的输出结果将作为下一层

的输入数据，在输出层得到最终结果。在(b)反馈型神经网络结构图中，节点根据各层其他节点的反馈不断地对自身输出进行调整，通过反复的迭代和调整最终使得输出结果趋向平衡。

人工神经网络具有良好的鲁棒性和容错能力，且分类质量高；在并行分布处理方面能力强。但也由于其复杂的网络拓扑结构，因此在计算中对计算机性能要求非常高。除此之外，由于神经网络的输出结果难以解释，因此其输出结果的可信度和可接受程度受到一定影响。

（3）集成分类。集成分类技术是训练多个分类器并对其预测结果进行组合从而建立一个分类器系统的方法。通过多个分类器的组合可以有效地克服某些分类器存在的缺陷，从而优化分类结果。常见的集成方法有自举汇聚法 Bagging、AdaBoost 和随机森林 RF。

Bagging 从原始数据集中随机选择得到 S 个与原始数据集大小相同的新数据集；然后将合适的学习算法分别作用于每个新数据集从而得到 S 个分类器。在对新数据集进行分类时，则使用这些分类器进行分类结果投票，投票最多的类别作为最终分类结果。Bagging 可以有效地减少模型的方差从而避免过拟合，有效提升分类结果的稳定性和准确性。

AdaBoost 是一种迭代算法。它根据各个分类器对样本权重和分类结果调整更新样本的权重；然后根据样本的权重重新选取获得新训练子集，针对新子集训练下一个基分类器[27]。AdaBoost 通过样本权值的改变控制样本的重要性，降低被正确分类样本的权重并提高被错分样本的权重[28]，从而提高被错分样本的重视率，最终提高分类器准确率。

另一种常用的集成分类方法是随机森林，下一小节将对随机森林进行详细的阐述与讲解。

3.2 随机森林算法的研究

由以上介绍可知，决策树分类算法存在受噪声影响大、过拟合的问题。为了克服这些缺点，美国科学院院士 Leo Breiman 利用了 Tin Karn Ho 等人提出的特征随机选择思想，并且综合 CART 决策树算法和 Bagging 集成学习，提出的一种新的组合分类器——随机森林。

随机森林（Random Forest，RF）是一个以决策树为基础的，具有良好分类性能的集成分类器算法。它在数据集的数据（行）和特征（列）的使用上进行随机选择；通过分裂规则生成很多决策树组成随机森林；最后新数据的分类结果按决策树投票结果而定。随机森林的分类器示意图 3-4 如下：

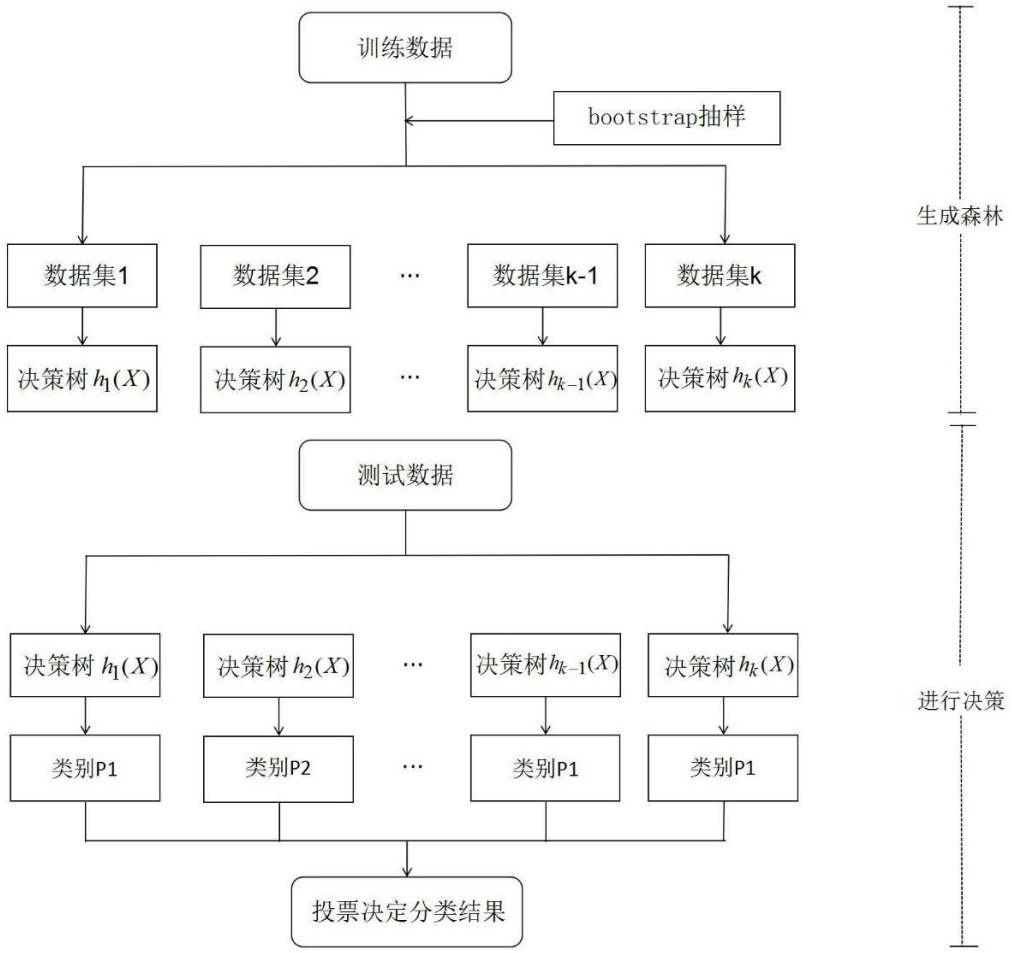


图 3-4 随机森林分类器示意图

假设原始训练集数据个数为 N，特征数量为 M，要构建的森林规模为 k，则随机森林的具体工作原理如下[29]：

1. 选取自助样本集。从原始训练样本集中有放回地随机抽取样本，形成一个新的子集——自助样本集。

2. 选取候选特征 。从步骤 1 中的自助样本集的 M 个特征中随机选择 m 个特征作为候选特征（ $m \le M$ ）。

3. 生成决策树。通过步骤 2 中选取的自助样本集的 m 个候选特征，计算每个节点的最佳分裂方式并对其进行分裂生长。对每一棵分类树的节点进行充分分裂直到每个叶子节点的不纯度达到规定要求，形成一棵决策树。

4. 继续构建新的决策树。重复以上步骤 1、2、3，根据不同的候选特征构建新的决策树，直到训练出 k 棵决策树为止。

5. 集成分类。将步骤 4 生成的多棵决策树组合成随机森林，使用随机森林对测试数据集进行分类预测；最终结果由决策树的分类结果投票决定，投票最多者则为最终分类结果。

随机森林是一项由多个弱分类器组成强分类器的分类方法。其实质是对决策树算法的一种改进：分别使用独立随机抽取的样品数据集建立一棵决策树，之后再根据建立的决策树集进行分类并投票产生最终结果。

设训练样本集为 $T=\{t_1,t_2,...,t_n\}$，生成的决策树集合为 $DT=\{d_{t_1},d_{t_2},...,d_{t_k}\}$，则随机森林的训练过程的算法描述如下所示[30]：

```
Algorithm RF_train(T,p,k,m){
    //训练 k 棵决策树
    for(int i=1;i<=k;i++)
    {
        //使用 bootstrap 重采样技术，有放回在训练集中抽取 p 个样本
        T′ = sample_withResample(T，p);
        //获取训练样本集的特征集
        Att= getAttributes(T′);
       //无放回地从特征集 Att 中抽取 m 个特征
       Att′ = sample_withoutResample(Att,m);
       //对 T′只保留 Att′中含有的特征
       T′′ = remainAttributes(T′,Att′);
       //构建决策树（设决策树数组为 DT）
       DT[i]=createDecisionTree(T′′);
    }
    return DT;
}
```

Breiman 指出，在训练过程中，从数据集中选取特征的个数 m 对随机森林的分类性能有比较大的影响[31]，因此实际应用中 m 值的选择是训练过程中的一个重要环节。

在随机森林的实现过程中，它通过对训练数据集的子集和特征的随机选择从而建立一棵决策树；再通过反复的随机选择生成 N 棵决策树最终形成“森林”。也正是这样的一个重要过程，使得随机森林具有以下的优点：

（1）通过对数据集的数据（行 ）和特征（列）的随机选择，使得随机森林在噪声较大的数据集中受影响较小，抗噪声能力较强。

（2）由于对训练样本和特征的随机性选择，使得随机森林的分类效果良好，克服了决策树容易产生的过拟合现象；即随机森林不容易陷入过拟合。

（3）随机森林在训练过程对特征进行随机选择，不用进行复杂的特征选择。

因此在处理成千上万高维度（即特征较多）数据时，有较快的速度。

（4）随机森林对数据集的适应性较强，特征的类型既可以是离散型，也可以是连续型，在分类过程中也无需对数据集规范化处理。

（5）在训练过程中，基于 OOB 误分率的增加量和基于分裂时 GINI 下降量，能够得到特征的重要性排序。

（6）在随机森林的构建过程中，每棵树的数据子集和特征的随机选择都是相互独立，没有影响的，因此随机森林容易做成并行化方法。

当然，随机森林也存在一些缺点[32]：

（1）在构建决策树时，对数据集子集和特征的随机选择，导致单棵决策树的预测效果很差。

（2）在随机森林建立过程中，挖掘者无法控制模型内部的运行，只能通过修改相应的参数来获取更好的结果。

（3）因为会对数据子集进行随机选择，因此当数据集中某一类的数据样本明显地少于其他类样本的数目即在处理非平衡数据集时随机森林的效果不佳。

（4）对于数据集含有多个不同级别的属性时，随机森林分类的结果受级别划分较多的属性影响较大，因此随机森林在这类数据集上的计算结果可信度不高。

3.3 随机森林填补非平衡数据集的改进

3.3.1 随机森林填补非均衡数据集的问题

在对中人口死亡数据集“死亡方式”这一属性缺失值填补中，已知的大多数死亡方式为非自杀，只有极少数的死亡方式为自杀，因此这便形成了类别不均衡数据。数据不均衡是分类任务中一个典型存在的问题。简而言之，即在数据集中，属不同类别的样本数目相差巨大，即一个类别的样本数总和远远多于另一个类别的样本数总和。在医疗信息中存在很多非平衡数据集。如在疾病诊断中，患有某种疾病的概率极小，因此绝大多数的诊疗者是健康的，只有极少数病人患有该疾病。又在医疗诈骗交易识别中，属于欺诈交易行为极少，即只有极少部分的交易属于欺诈交易，从而造成正常交易与欺诈交易之间数据量相差甚大。这些都是医疗领域中常见的不均衡问题。

上一小节中也提及到随机森林在处理类别不均衡数据时的分类较差。一方面是由于每棵决策树的训练子集是通过 bootstrap 重采样技术随机挑选的，因此少数类样本数据比较不容易选中，加剧了类别不均衡现象。另一方面，由于数据集中少数类别较少，因此构建的决策树也没能很好的体现少数类样本的特点，最终导出分类效果不佳。在数据挖掘领域，如何对非平衡数据集进行处理成为十大挑战

难点之一。本文在对数据集“死亡方式”这一属性数据进行缺失值填补也同样受到非平衡数据集问题的影响，因此在使用随机森林进行类别填补前需对数据集进行平衡改善。下面将对处理非平衡数据集的方法和分类性能评价指标做简要阐述。

1、改善非平衡数据集方法

处理非平衡数据集问题的方法有改进分类算法和对数据集进行改造。改进分类算法主要是根据数据集的自身特点对分类的算法进行改进优化，从而减少类别不均衡所造成的错误分类。对于这一类方法，数据集本身不会发生变化，主要有单类学习和代价敏感。

单类学习在类别不均衡严重的情况下使用，它仅对少数类的数据进行训练。在对新样本进行分类时，分析其与少数类样本的相似性从而判断其类别。代价敏感学习则是一个 boosting 算法[33]。它设定把少数类数据错分所产生的代价要远大于多数类数据错分的代价，从而有效提高少数类样本在分类器中的识别率[34]。

对数据集进行改造方法的实现则是对数据集进行增减。它的基本思想是增加或者减少不平衡数据集样本的个数，从而改变数据集的分布来消除或减小数据的不平衡。主要有欠采样技术和过采样技术。

欠采样技术将多数类样本进行适当删减，从而使数据集趋于平衡。常用的方法是随机欠采样技术，它通过随机选取的方式，将部分多数类样本删除从而减小其规模，其存在的缺点是伴随着多数类样本的删除其携带的某些重要信息也会随之丢失，从而造成分类器分类性能的下降。

过采样技术则与欠采样技术相对，该方法通过增加少数类样本数量最终达到改善非平衡数据集的目的。最简单的一种方法是随机过采样。它通过随机选择的方式，将部分少数类样本复制添加到原始数据集中从而提高少数类样本的比例；这种方法的缺点是添加的少数类样本与原始数据集的部分样本重合，可能导致过拟合现象的发生。

2、非平衡数据集性能评价指标

分类算法很多，不同的分类算法有自身的特点，在不同的数据集上表现出的效果也自然不同。对非平衡数据集，通常使用分类精度、ROC 曲线[35]和 F1-Measure 这三类指标对分类效果进行评价。

表 3-1 模型评价术语解析表

实际类别	预测类别			
		Yes	No	总计
	Yes	TP	FN	P（实际为 Yes）

续表 3-1 模型评价术语解析表

	No	FP	TN	N（实际为 No）
	总计	P′（被分为 Yes）	P′（被分为 No）	P+N

如表所示，假设分类目标只有两类，记为正类和负类。则在实际分类中，会存在以下四种情况：真正类(True Postive TP)、假负类(False Negative FN)、假正类(False Postive FP)、真负类(True Negative TN)。在此基础上，分类指标的计算如下。

（1）分类精度（ Accuracy ）。这个指标表示被分对的样本数除以所有的样本数，它用来衡量随机森林对测试数据集的总体分类情况，计算公式如下：

$$accuracy=(TP+TN)/(P+N) \tag{3-1}$$

一般来说，分类精度高，则分类效果好。

（2）ROC 曲线(Receiver Operating Characteristic)。ROC曲线是反映 FPR 和 TPR 的综合指标。横纵坐标分别为 FRP 和 TPR：

$$FPR=FP/(FP+TN) \tag{3-2}$$

$$TPR=TP/(TP+FN) \tag{3-3}$$

FPR 越大，预测正类中实际负类越多；TPR 越大，预测正类中实际正类越多。分类性能好的 ROC 曲线是在 FPR 取值较小时有较大的 TPR。AUC 面积则为 ROC 曲线之下的面积。总而言之，当 ROC 曲线越倾向于左上方，AUC 面积越大，算法的分类处理能力就越强。

（3）F1-Measure。F1-Measure 也是非平衡数据集的一个重要性能评价指标，它在自然语言处理和信息检索中的到广泛应用。F1-Measure 的计算方法如下：

$$Precision=TP/(TP+FP) \tag{3-4}$$

$$\mathrm{Re}call=TP/(TP+FN) \tag{3-5}$$

$$F1=2*\mathrm{Re}call*\mathrm{Pr}ecision/(\mathrm{Re}call+\mathrm{Pr}ecision) \tag{3-6}$$

对于分类的结果， $Precision$ 和 $\mathrm{Re}call$ 都是越高越好，但在实际应用中，两者可能存在着互相矛盾。因此使用 $F1-Measure$ 指标对这两者进行综合评价[36]。同样的，$F1-Measure$ 越高，说明分类效果越好。

3.3.2 SMOTE 算法

为了克服随机过采样技术导致的过拟合这一问题，Chawla 等人于 2002 年在人工智能杂志上提出了一种新型的过采样技术 SMOTE 算法来合成增加少数类样本。

SMOTE 是一种改善非平衡数据集的过采样技术。这种方法与 KNN 的理论基础很相似，即少数类样本的邻近样本其类别与少数类相同。基于这样的理论，该方法为：在少数类样本的邻近区域合成新样本，并将其作为新增少数类样本添加到原始数据集中。SMOTE 通过提高少数类样本在非平衡数据集中的比例，有效扩大了分类决策的区域。同时由于合成新样本与少数类样本的差异性，因此可以防止分类器出现过度拟合。

SMOTE 算法的具体方法简单地概括为：对于每个少数类样本 x_i 确定其 k 个距离最近的少数类样本记为集合 S；若设定向上采样倍率为 m，则从集合 S 中随机抽取 m 个样本（$k > m$）记为集合 M；然后将少数类样本 x_i 与集合 M 的每个样本分别进行随机线性插值，合成新的少数类样本并添加到数据集中；最后判断新数据集的非平衡率（即少数类数目占全体样本数目的比例），若还是过小则重复以上步骤否则停止。构造新样本的公式如下：

$$p_{ij} = x_i + rand(0,1) * (y_{ij} - x_i) \tag{3-7}$$

其中，$x_i(i = 1,2,...,n)$ 代表一个少数类样本，n 表示少数类样本的总数量；$y_{ij}(j = 1,2,...,m)$ 为与样本 x_i 相邻的 m 个近邻样本；$p_{ij}(j = 1,2,...,m)$ 表示合成的新少数类样本；$rand(0,1)$ 则表示（0，1）之间的一个随机数。

下面对 SMOTE 算法做一个简单的实例演示。如图 3-5：

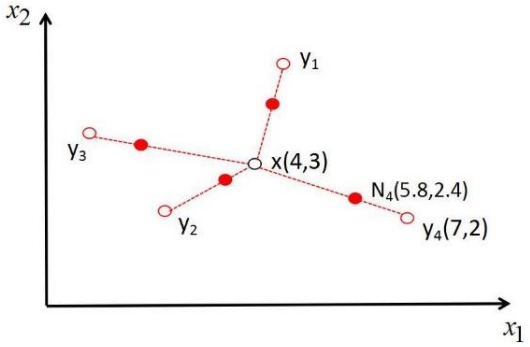


图 3-5 SMOTE 算法简单实例图解

对于样本点 x，选取了它的 4 个最近邻数据样本，分别为 y_1, y_2, y_3, y_4，现在对 x 和近邻点 y_4 合成新样本点 N_4。则根据公式，当随机数 rand(0,1)为 0.6 时可得到：

$$\begin{aligned} N_4 &= x + rand(0,1) * (y_4 - x) \\ &= (4,3) + 0.6 * ((7,2) - (4,3)) \\ &= (4,3) + (1.8, -0.6) \\ &= (5.8, 2.4) \end{aligned} \tag{3-8}$$

因此，新合成的少数类样本点为 $N_4(5.8, 2.4)$。x 与 y_1, y_2, y_3 的合成的新样本点也可

通过以上方法得到。

SMOTE 算法的程序流程图如图 3-6：

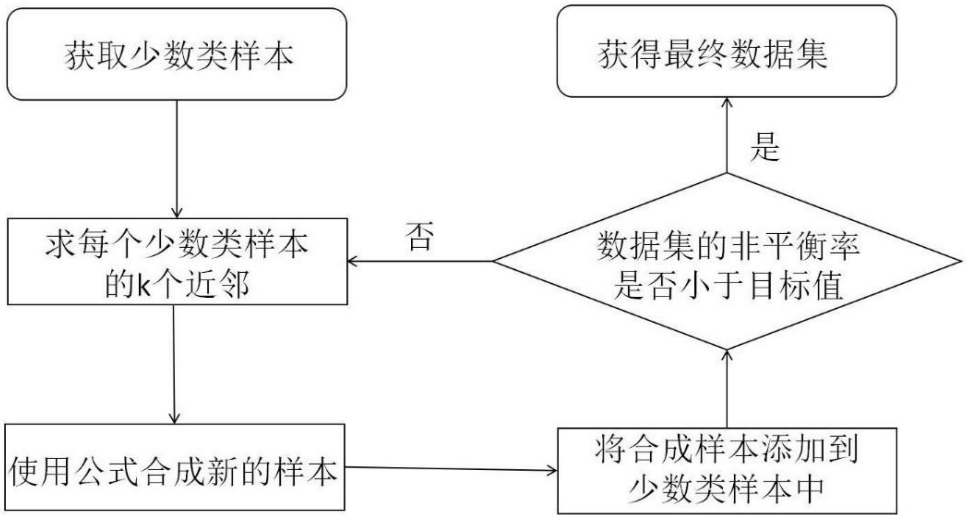


图 3-6 SMOTE 算法流程图

不同于随机过采样技术通过简单复制少数类样本来改善非平衡数据集，SMOTE 算法有效地避免了随机过采样技术在添加新样本时的局限性和盲目性。它利用线性插值的基本数学理论合成的新样本，这些新样本拥有了少数类样本的某些特性，在不会与原数据集中的样本重合的情况下增加了少数类样本的数量，扩大了分类决策的区域的同时又改善了数据集的非平衡性。

尽管 SMOTE 算法在一定程度上缓解了随机过采样技术的局限性和盲目性，SMOTE 也依然存在着问题。

首先，SMOTE 算法在近邻数 k 的选择上没有可依据的标准，具有一定的盲目性。在 SMOTE 算法中，k 的近邻时没有太大限制的；但当数据集中存在噪声时，合适的 k 值有可能不会使得某些新样本也成为噪声。如图 3-7 所示：

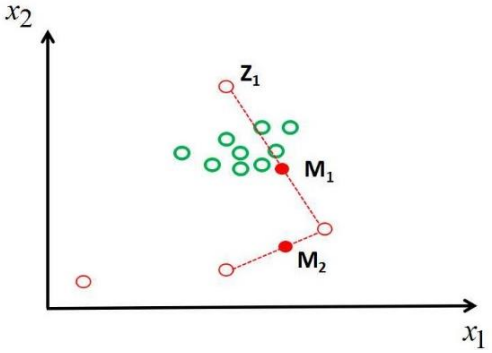


图 3-7 SMOTE 算法在噪声样本中合成新样本

图中可以看出，少数类红类合成的新样本 M1 和 M2 中，M2 是比较合理的。但因为噪声样本 Z_1 的存在，使得合成的新样本 M1 也成为了数据集的噪声样本。

这样的数据不仅无法提升分类的准确度，反而增加了数据集的噪声，最终影响分类器的分类效果。

其次，若一个少数类样本处于少数类数据集的边缘，则其选择的近邻也很有可能是边缘样本，从而导致生成的新样本也成为边缘样本。如图 3-8 所示：

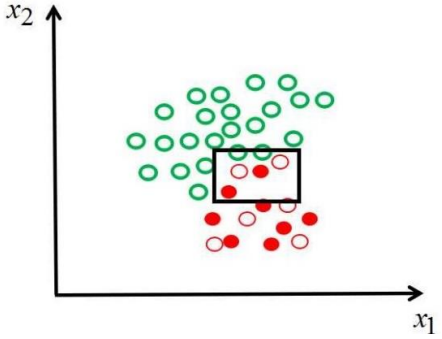


图 3-8 SMOTE 算法在边缘样本中合成新样本

图中可以看出，因为两个少数类样本是处于样本边缘的，因此通过 SMOTE 算法合成的新样本也处于样本边缘。随着合成新样本数据的增多，边缘化样本比例不断增加，使得样本类别之间的界限越来越模糊，进一步增加了分类的难度。

3.3.3 基于重心的 SMOTE 算法的改进

基于上小节中对 SMOTE 算法的分析可以知道，由于需选取样本的最近邻数 k 值，当数据集存在噪声样本时，不适当的最近邻数 k 值会导致合成的新样本很可能成为数据集的噪声。除此之外，对于边缘性样本而言，其近邻也常常是边缘性样本，因此合成的新样本也最终成为了边缘性样本，模糊了少数类和多数类的界限，给下一步的分类造成了严重的影响。因此，为了克服以上问题，本小节对 SMOTE 算法进行改进提出一种基于重心的 SMOTE 算法。

基于重心的 SMOTE 算法的依据有以下两点理论：

（1）根据物理学和几何学的原理，属于同一个类的样本应该有一个共同的重心[37]。这个重心往往表征了这类样本的独有特性。如下图 3-9 所示：

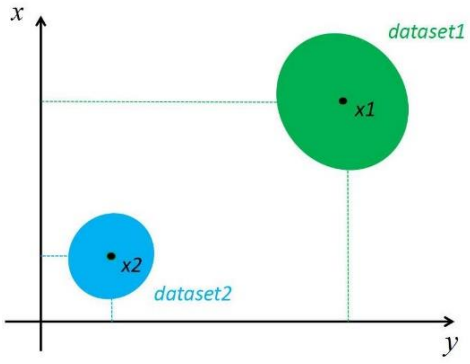


图 3-9 数据集的重心

x1 和 *x2* 分别是 dataset1 和 dataset2 的重心。可以大致看出，重心 *x1* 的位置特点是具有较大的 *x* 和 *y* 取值；而重心 *x2* 的位置特点是 *x* 和 *y* 都取值较小，符合现实情况。因此，虽然数据集的样本多且呈现出各种各样的特性，但数据集的重心可以大致反映整个数据集的整体分布情况及数据集的共有特点。

（2）由样本重心与两个同类样本建立的区域应该是该类样本的共有区域，即在这个区域空间内的样本都应该是属于同一类。

因此，基于以上两点理论，若在共同区域内合成新的样本，那么这些新的样本既跟原始样本有差异，但同时又可以保留住同类样本的一些固有特性；从而有效的克服了 SMOTE 算法在噪声样本中因为近邻的选择而合成噪声新样本的问题。又因为在建立共同区域时是随机选择两个同类样本和重心点的，因此同时选中两个边缘样本的概率大大降低，有效克服了由边缘性样本造成的模糊类别界限的问题。改进的 SMOTE 算法通过两次计算区域的重心，从而使得新合成的样本具有一定的区域性，也更能集成少数类样本的某些共同特性。

假设训练数据集少数类样本的总数为 n，则其集合表示为 X: $X=\{X_1,X_2,...,X_n\}$；若每个样本具有 m 个属性，则每个样本表示为 $X_i:(x_{i1},x_{i2},...x_{im})$。多数类样本集合表示为 Y: $Y=\{Y_1,Y_2,...,Y_l\}$，则基于重心的 SMOTE 算法的具体步骤如下：

Step1. 计算少数类样本的重心点，记为 X_g。这里采用向量和欧式距离的计算方式和得到少数类样本的重心点：

$$X_g=\frac{1}{n}(\sum_{i=1}^{n}x_{i1},\sum_{i=1}^{n}x_{i2},...,\sum_{i=1}^{n}x_{im}) \tag{3-9}$$

根据公式可以计算出少数类样本的重心点 X_g。

Step2. 构建一个少数类小区域的重心点，记为 X_c。从少数类样本集合 X 中随机选取两个样本，分别记为 x_{r_1},x_{r_2}。通过三个样本 X_g，X_{r_1}，X_{r_2} 求取该小区域的重心点 X_c。公式如下：

$$X_c=\frac{1}{3}((x_{g1}+x_{r_11}+x_{r_21}),(x_{g2}+x_{r_12}+x_{r_22}),...,(x_{gm}+x_{r_1m}+x_{r_2m})) \tag{3-10}$$

通过这个公式可以求取到少数类小区域的重心，从而使得新生成的样本有一个靠近的区域中心方向。

Step3. 合成新的样本 p_i。为了保证新合成的样本向小区域中心 X_c 靠近，因此对 SMOTE 算法的新样本合成公式进行了改进。公式如下：

$$p_i=X_i+rand(0,1)*(X_c-X_i) \tag{3-11}$$

其中，$X_i(i=r_1,r_2)$ 为步骤 2 中随机选择的两个少数类样本；$p_i(i=r_1,r_2)$ 为合成的新

样本；rand(0,1)取值同上为(0,1)之间的一个随机数。将新生成的样本 p_i 放入到数据集合 X_{new} 中。

Step4.计算非平衡率。计算数据集的非平衡率 R，公式如下：

$$R=\frac{count(X)+count(X_{new})}{\text{count}(X)+count(X_{new})+count(Y)} \tag{3-12}$$

如果非平衡率小于目标值则继续重复步骤 2,3,4 以获得更多的合成样本；若平衡率达到目标值则合成新样本结束，最后将合成的数据集 X_{new} 放入原始数据集中作为随机森林分类算法的训练数据。

基于重心的 SMOTE 算法的流程图则如下图 3-10 所示：

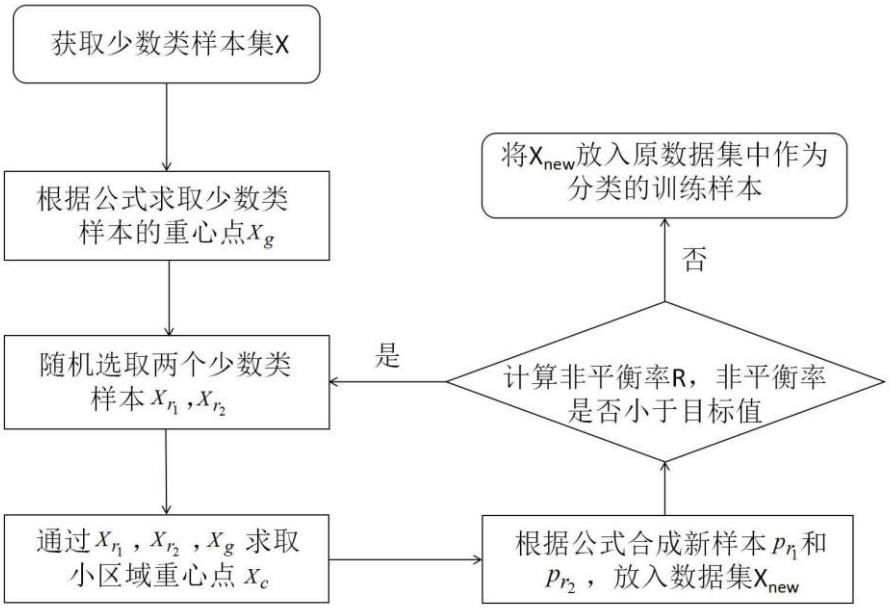


图 3-10 基于重心的 SMOTE 算法流程图

从流程图可以看出，该方法通过构建重心点与两个同类样本的共同区域，从而在这个区域中合成具有相似特征的新样本，最终达到改善非平衡数据集的目的。

如图 3-11 为算法的具体过程：

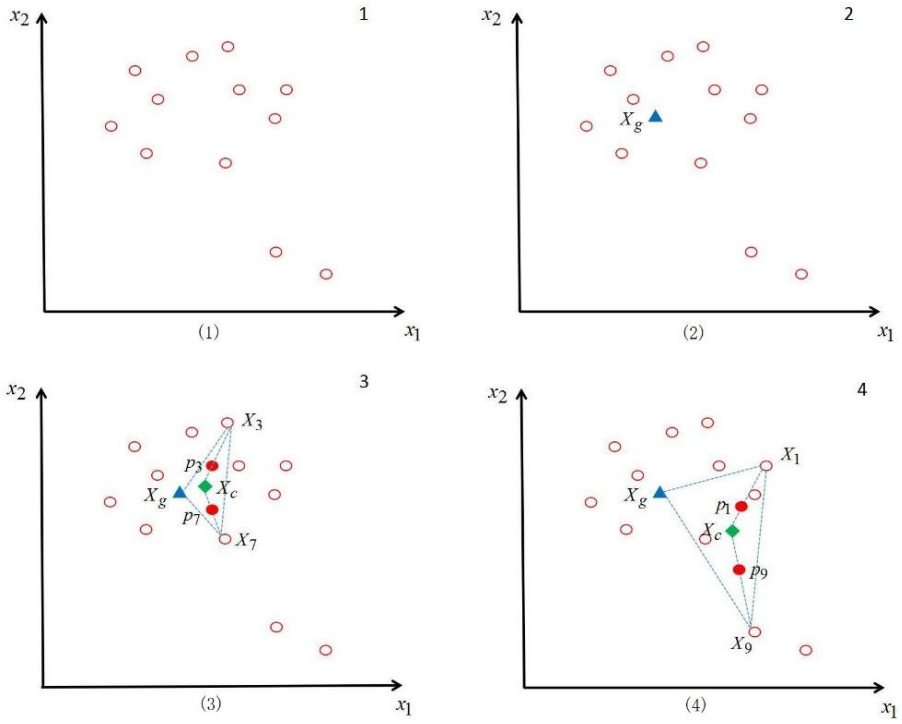


图 3-11 基于重心的 SMOTE 算法实例图解

首先，（1）图表示少数类样本的整体情况，使用改进的 SMOTE 算法增加少数类新样本从而改善数据集。根据计算公式得到少数类样本的重心点 X_g， X_g 就是图（2）中的蓝色三角形。之后选取到随机样本点 X_3、X_7，并通过相关公式计算其与 X_g 构成区域的中心 X_c， X_c 则是由（3）图中的绿色四边形表示。最后利用公式计算出新的合成样本 p_3， p_7；从图中可以看出新合成的样本点在三角区域 $X_3X_7X_g$ 中，跟其他同类样本具有较高的相似度。而如（4）图所示，即使在噪声点 X_9 的干扰下，新合成的样本点 p_9 也是尽可能地向少数类数据集合靠近的，较多地继承了少数类样本的共同特点。因此改进的 SMOTE 算法从一定程度上克服了噪声数据点的干扰。

而处理边缘样本的情况则如图 3-12 所示：

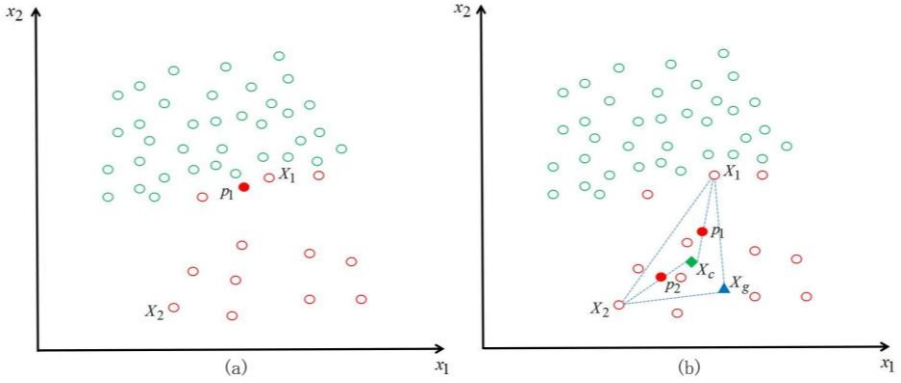


图 3-12 基于重心的 SMOTE 算法处理边缘样本的对比分析图

在(a)图中，使用传统的 SMOTE 算法，通过边缘样本 x_1 合成的新样本为 p_1，可以清晰看到新合成的样本 p_1 同样为少数类数据集红类的边缘样本，这在一定程度上模糊了少数类和多数类样本的分界线，增加了分类的难度。而(b)图则是使用改进的 SMOTE 算法合成的新样本 p_1，从图中可以看出，新合成的样本 p_1 在向样本的中心靠近，已经具有了少数类样本红类的某些共同特性，与多数类样本绿类存在明显的差异性。通过这种方法良好的区分了少数类与多数类，克服了使用 SMOTE 算法在处理边缘样本时的缺陷。

3.4 实验结果

3.4.1 非平衡数据集的实验

在这一小节，我们获取了三组类别不均衡的纯数值数据集，使用传统的 SMOTE 算法和基于重心的改进 SMOTE 算法对数据集进行平衡改造，再使用随机森林进行分类预测。通过两者的结果对比，对改进的 SMOTE 算法的效果进行验证。

在 UCI 机器学习数据库中我们收集到实验中所使用的的数据集，数据集可以通过网址 http://archive.ics.uci.edu/ml/ 下载得到。实验中的三组数据集来自不同的领域，它们的非平衡率、属性类型和属性数量均不相同。这三组数据集分别为 Glass、Adult 和 Haberman。Glass 数据集是根据玻璃的各种属性判断是否为犯罪常用玻璃类别；Adult 数据集是根据成年人的个人状况判断是否年薪在 50K 以上；Haberman 数据集则是根据乳腺癌患者的情况对手术后的存活或死亡情况进行判断。表 3-2 对这三组数据集的信息进行了详细描述：

表 3-2 三组数据集的详细描述

数据集	属性个数	样本数	少数类样本数	多数类样本数	非平衡率
Glass	10	214	29	185	13.6%
Adult	14	48842	12732	36110	26.1%
Haberman	3	306	81	225	26.5%

由表中可知，这三组数据的非平衡率都较低，因此在少数类样本较少的情况下，使用随机森林对样本数据集进行分类预测难度较大。因此这里采用了传统的 SMOTE 算法和改进的基于重心的 SMOTE 算法对数据集的非平衡性进行改善；之后通过随机森林对数据集进行分类预测。并使用分类精度、ROC 曲线（AUC 面积）以及 F1-Measure 评价指标对实验结果进行对比分析：

（1）分类精度 Accuracy：

表 3-3 分类精度结果对比表

数据集	Glass	Adult	Haberman
原始数据集	0.6542	0.6669	0.6937
传统 SMOTE 处理	0.7319	0.7523	0.7875
改进的 SMOTE 处理	0.7748	0.7842	0.8273

（2）AUC 面积：

表 3-4 AUC 面积结果对比表

数据集	Glass	Adult	Haberman
原始数据集	0.6812	0.7043	0.7137
传统 SMOTE 处理	0.7511	0.7685	0.7963
改进的 SMOTE 处理	0.7885	0.7933	0.8273

（3）F1-Measure：

表 3-5 F1-Measure 结果对比表

数据集	Glass	Adult	Haberman
原始数据集	0.7041	0.6829	0.7337
传统 SMOTE 处理	0.7897	0.8185	0.8212
改进的 SMOTE 处理	0.8159	0.8433	0.8533

从表 3-4，3-5，3-6 可以看出：

（1）因为这三组数据的非平衡率都较低，即少数类占整体数据样本的比例较低；因此当对这三组数据集使用随机森林算法直接进行分类时，虽然不同的数据集分类效果有波动，但可以明显看出，其整体的分类性能都不高。

（2）使用传统的 SMOTE 算法合成少数类样本提高数据集非平衡率，然后再使用随机森林算法对数据集进行分类预测。三组数据集的分类性能都得到了一定的提高，平均提高了 10%。这一结果表明，当使用随机森林算法对数据集进行分类时，数据的不平衡会在一定程度上影响分类的结果。

（3）而使用改进的 SMOTE 算法对数据集进行平衡改造，再使用随机森林算法对数据集进行分类预测。三组数据的分类性能在使用传统 SMOTE 算法上又得到了进一步的提高，平均提高了 3%-5%。这一结果表明，改进后的 SMOTE 算法合成的新样本质量更优。

3.4.2 随机森林填补人口死亡方式缺失值

上一小节，我们对 UCI 机器学习数据库中的三个非平衡数据集进行了实验，实验结果表明：使用传统的 SMOTE 算法改善非均衡数据集能够提高随机森林的的分类性能；同时改进后的 SMOTE 算法合成的新样本比传统的 SMOTE 算法质量更优，随机森林的分类效果更佳。

这里使用传统的 SMOTE 算法和改进后的算法改造“人口死亡”数据集的非平衡性；之后再使用随机森林对数据集的人口“死亡方式”缺失值进行分类预测并进行填补。数据集取自 https://www.kaggle.com/cdc/mortality，数据集的具体情况如下：

表 3-6 “人口死亡”数据集详细信息

数据集	属性个数	样本数	少数类样本数	多数类样本数	非平衡率
Death	22	38601	5983	32618	15.5%

通过数据的相关性分析删除了一些属性特征，最终数据集做分类处理时保留的数据特征为 age，sex，nativePlace，disability，education，occupation，activityLove，mentalDisease。在实验中调整随机森林算法进行分类时使用的参数，最终确定使用 n_estimators=1200，max_features=6 即决策树棵树为 1000 棵，最大特征数为 6 时随机森林的分类结果最佳。在此基础上，对原始数据集使用传统 SMOTE 算法和改进 SMOTE 算法进行平衡改造，实验对比结果如下：

（1）分类精度 Accuracy：

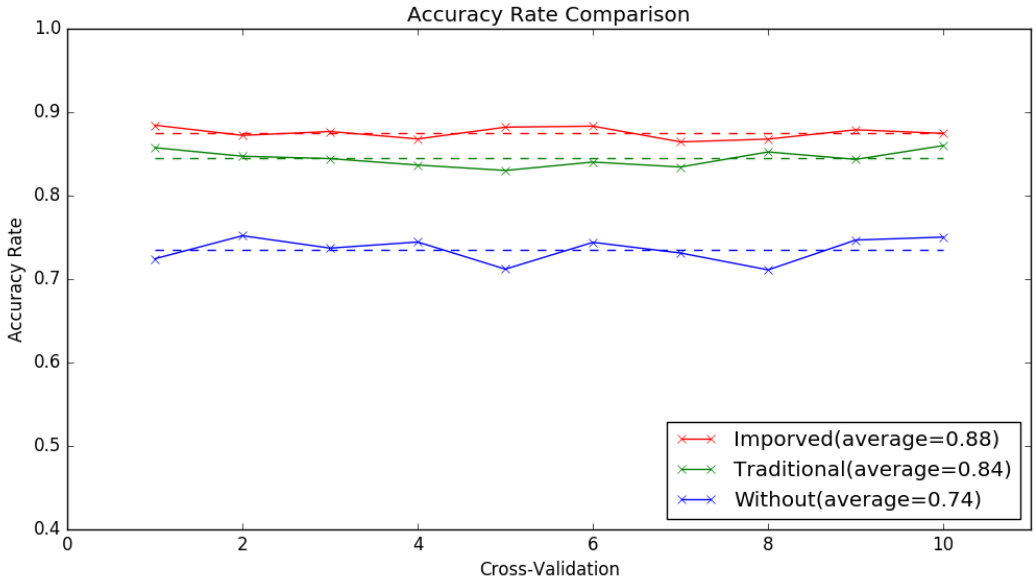


图 3-13 cv=10 时的分类精度结果对比

可以清晰看到，使用改进后的 SMOTE 算法的平均分类精度明显高于传统的

SMOTE 算法 4 个百分点；而未使用 SMOTE 算法改善原始数据集时的分类精度则大大低于传统 SMOTE 算法。

（2）ROC 曲线：

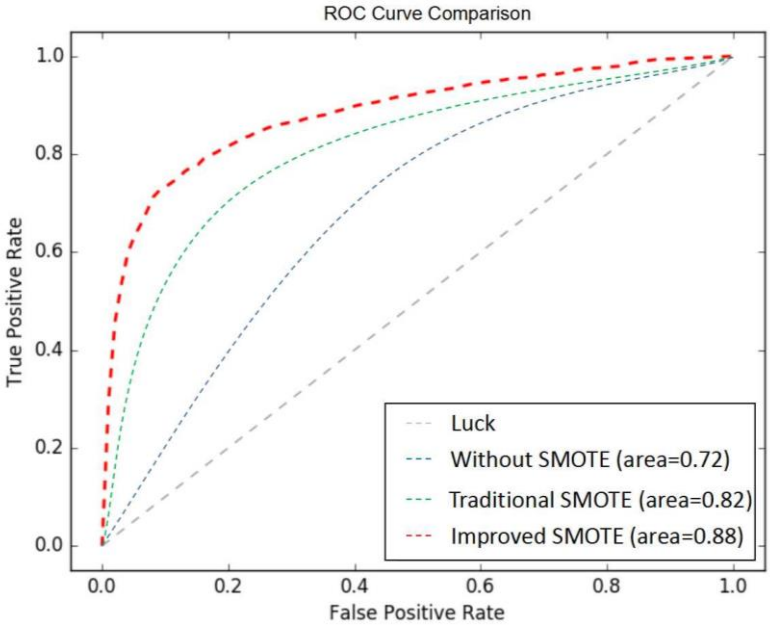


图 3-14 ROC 曲线结果对比

图中可以看到，使用改进的 SMOTE 算法改善原始数据集时的 AUC 面积最大，达到 0.88，比传统的 SMOTE 算法高出 0.04，而使用随机森林对原始数据集进行分类其 AUC 面积则只有 0.72，分类性能较差。

（3）F1-Measure：

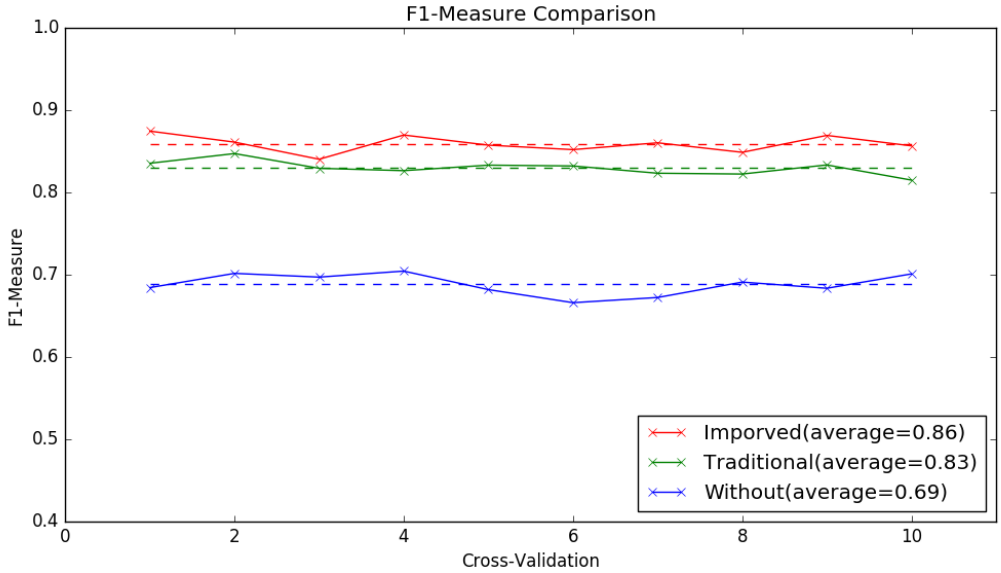


图 3-15 cv=10 时的 F1-Measure 结果对比

图中可以看出，使用基于重心的 SMOTE 算法对数据集进行平衡改善并使用随机森

林分类时，其 F1-Measure 是最高的，分类效果明显优于其他两种。

从上面的分类精度、ROC 曲线、F1-Measure 的实验结果可以看出，使用随机森林直接对原始数据集进行分类，其各项评价指标都只在 60%-70%左右，分类性能较低。当使用传统 SMOTE 算法对原始数据集进行平衡改造时，随机森林的分类性能得到一定的提高，大致提高了 10-15%。而使用改进后的 SMOTE 算法对原始数据进行平衡改造后，随机森林的分类性能又得到进一步的提高，平均提高了 4%。这表明，改进后的 SMOTE 算法可以克服传统 SMOTE 算法处理噪声和边缘性样本的问题。因此，随机森林处理在处理非平衡数据集时的分类效果有显著的提高。

3.5 本章小结

本章主要研究了“人口死亡”数据集中“死亡方式”属性缺失值的现代填补方法。归根结底，缺失值填补就是对训练数据集建立分类模型从而对缺失部分有效预测。因此，本章先对几种常用的分类方法进行了简要阐述，之后对随机森林分类进行了深入的研究。由于数据集的非平衡性，随机森林算法的分类性能较差，因此进一步研究了随机森林处理非平衡均衡数据集的解决方法。第三小节主要对改善数据集平衡性的 SMOTE 算法进行了研究，同时对 SMOTE 算法提出创新性改进并在第四小节进行实验对比分析。实验结果表明：通过使用改进的 SMOTE 算法改善非平衡数据集，可以有效提升随机森林算法的分类效果。

第四章 癫痫病脑电波频域信号降维处理技术的研究与改进

4.1 降维技术的研究

4.1.1 降维的必要性

高维数据集无处不在：图像、生物医学数据、声音信号、金融数据、光谱等[38]，它们很多时候都是通过几百个乃至成千上万个属性来描述的。例如在人脸识别中，对于分辨率仅为 256*256 人脸图像采用行堆叠的方式将其转换，则每张人脸需要通过 65536 维数据来表示。随着数据量的不断增加和数据信息维度的指数型提升，如何从繁杂的原始数据中较为高效的挖掘出有用的信息是研究的一个重要方向。

本章研究的脑电波（Electroencephalo-graph，EEG）频域信号具有高维度的特性。脑电波是由大量脑神经细胞组合而成，它反映的是在高度相关状态下，电活动在头皮上的总体效应。通过对脑电波的检测，可以较为直观的反映“活”的脑组织的功能状态。科学家在此领域近几十年的研究表明，脑电波的研究在神经系统疾病如癫痫病的诊断方面一直发挥着重大作用。本文处理的脑电波数据是医院在采样频率为 5000HZ 下获取的一段时间的脑电波信号，这意味着一位病患一个周期的脑电波数据的维度便高达 5000 维。这些脑电波信号不仅呈现着复杂的非线性结构[39]，并且具有信号数据噪声干扰性强、维度高的特点。庞大复杂的高维度脑电波数据不仅会大大的降低数据挖掘的效率，同时在一定程度上影响挖掘的结果。因此对脑电波数据进行预处理是非常有必要的；而对高维度的脑电波频域信号采取合适的降维处理便是本小节研究的一个重点。

4.1.2 降维相关算法的研究

简单来说，降维技术的主要思想是：给定一个高维数据集 $X=\{x_1,x_2,\cdots,x_n\}\subset R^D$，在一定差错容忍范围内，找到一个低维度的数据集 $Y=\{y_1,y_2,\cdots,y_n\}\subset R^d(d\le D)$ 来代替高维数据集 X，从而降低数据挖掘的难度。降维技术的具体应用包括数据压缩、模式分类与识别、数据可视化和多媒体信息检索[40]等。

根据数据之间的相互关系，可以将降维技术划分为线性降维和非线性降维。线性降维技术通常假定数据集来自一个全局线性的高维空间，即数据的各变量是相互独立的，通过线性降维把数据投影到低维线性子空间。较为常用的算法包含以下三种。

（1）主成分分析（ Principal Components Analysis，PCA ）。PCA 是使用最广

泛的经典线性降维方法之一[41]。PCA 方法的主要思想是寻找一个新的坐标系 L，将具有高维度的特征空间数据沿着 L 映射到低维的特征空间中，并尽可能保留其原始数据的主要信息。在投影后的新坐标系下，投影后的投影值分布越散,则表明保留的原始信息越多。如下图 4-1 所示：

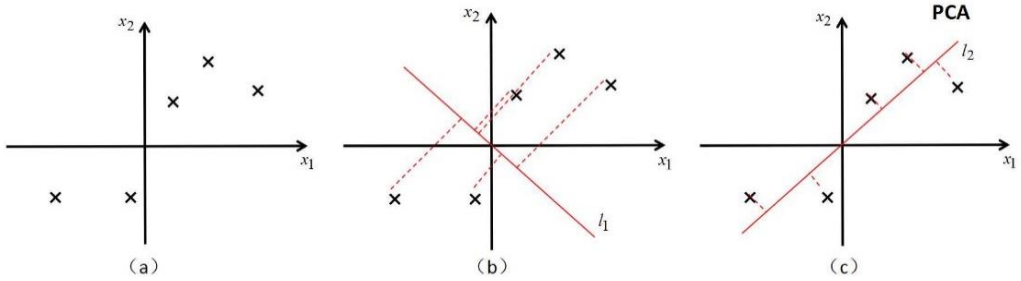


图 4-1 PCA 降维

要对（a）图中的这个二维的数据集进行降维，则选择的投影坐标可以如图（b）图（c）；在（b）图中，原数据投影的低维数据之间的间距比较密集；而在（c）图的 PCA 降维中，参考投影向量 l_2 使投影后的新数据分布比较分散，即方差较大，则表明新数据集更多的保留了数据集的原始信息。

（2）线性鉴别分析（Linear Discriminant Analysis，LDA）。它的基本思想是将高维数据的特征空间映射到最佳鉴别矢量空间[42]，通过这样的方式不仅可以降低特征空间维度而且还易于抽取原始数据集的类别信息。经过投影后的样本子空间，使得同类样本距离越近，不同类样本距离越远。如下图 4-2 所示：

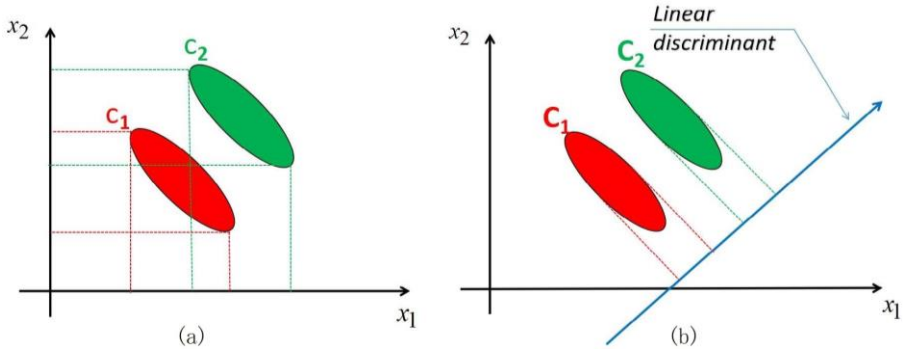


图 4-2 LDA 降维

要将图中的绿色类和红色类从二维降维到一维。如(a)图所示的投影方式，不同类别之间出现信息重复，不利于分类。而(b)图采用 LDA 方法后，红色类和绿色类在映射后距离最大，而且同类数据分布明显较为集中紧凑。LDA 将数据的类别信息作为主要参考对象，使得降维后属于不同类别的数据点距离分散，而同类别的数据点分布集中紧凑。由于 LDA 是有监督的降维算法，因此常用于分类挖掘。

（3）多维尺度方法（Multidimensional Scaling，MDS）。MDS 更多的保持了

原始数据集中样本的差异性，它的主要思想是：如果原数据集合原本相近的点在降维后依然相互靠近；反之亦然。MDS 利用数据点间的距离或顺序信息从而达到降维目的。

线性降维方法因简单而得到广泛的应用。但现实世界数据集更多地呈现出非线性结构，线性降维方法难以揭示其潜在本质信息。因此，人们提出了非线性降维方法来处理高维空间中的复杂数据点[44]。具有代表性的非线性降维算法主要包括：等距映射 ISOMAP、拉普拉斯特征映射 LE 和局部线性嵌入 LLE。

（1）等距映射（Isometric Mapping，ISOMAP）[45]。ISOMAP 在现有的 MDS 的基础上进行改善。在计算样本间距离时 MDS 采用的是欧式距离，而 ISOMAP 更多地考虑数据间的内在几何性质[46]，采用测地距离作为样本距离评估标准。

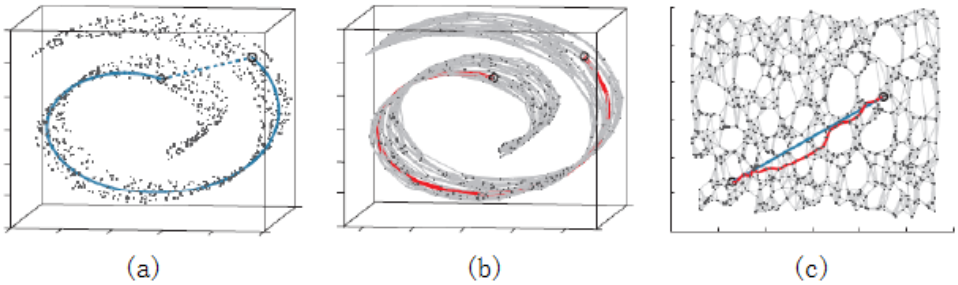


图 4-3 ISOMAP 降维

图 4-3 中，(a)图蓝色虚线表示的是样本点间的欧式距离；而在(b)图，样本点间的距离则由测地距离表示，在流行结构中更好的反映了点与点之间的真实距离，保留了原始数据集的真实形态[47]。ISOMAP 是一种全局优化方法，它的降维结果良好地保持了样本点间的拓扑结构。

（2）拉普拉斯特征映射（Laplacian Eigenmap，LE）。LE 利用局部信息构建数据间的关系[48]，它的直观思想是希望通过互相间靠近的点在降维后的空间中也尽可能的相近，从而保留数据内在的流形结构，因此在分类问题上有比较好的降维结果。如图 4-4：

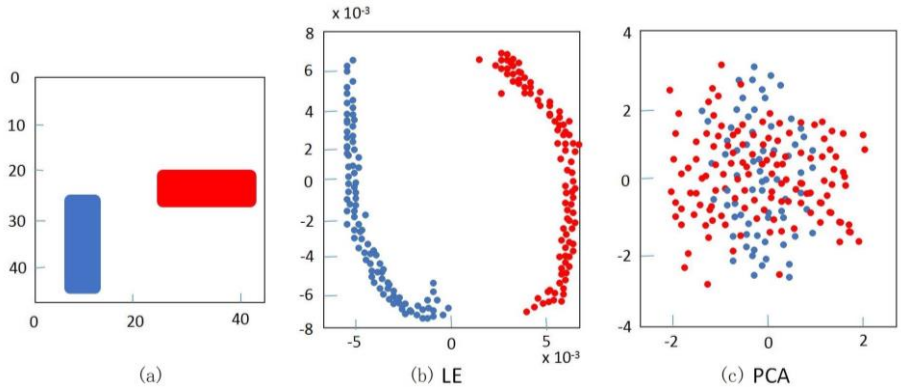


图 4-4 LE 降维实例

如上图所示，(a)图表示有两类数据点（数据是图片），使用LE和PCA方法将数据降到二维时的效果如图(b)、(c)所示。可以清楚地看到，在此分类问题上，LE的结果明显优于PCA。

（3）局部线性嵌入（Locally Linear Embedding LLE）。它是一种局部的优化算法，更多地考虑近邻样本，具有较好的数据局部保持性。本章下一小节将对该算法进行详细讲解。

4.2 局部线性嵌入算法的研究

局部线性嵌入（Locally Linear Embedding LLI）是Roweis等研究者针对流形结构数据提出的一种降维算法[49]。在现实生活的实际应用中，LLE 算法的降维效果比线性降维算法的降维效果更为优异，适用范围也更广。

LLE 算法是建立在简单的几何直觉基础上的，即在高维数据特征空间中的邻近点在低维映射空间中也依然保持较近的距离，从而将高维空间数据集映射到低维空间中[50]。根据这一假设，LLE 算法的主要思想为：为高维特征空间中的每个数据样本点寻找一个邻近区域，并且给该区域中的每个样本点赋予一个权重（这些权重象征着样本点的局部空间特征）；之后再求取局部重建权值矩阵用于描述样本点间的拓扑关系；最后根据这些权重寻找高维数据的最佳低维映射表示。如下图 4-5 描述了算法的具体实现过程：

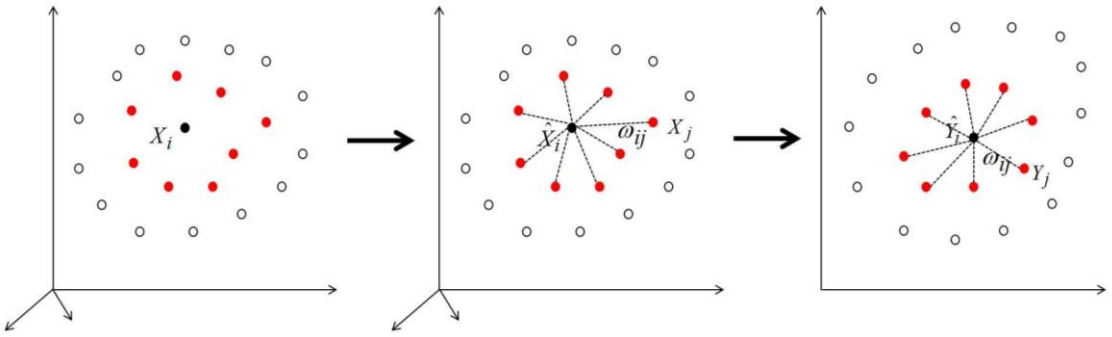


图 4-5 局部线性嵌入算法的实现过程图解

为方便叙述，定义如下符号：D 表示原始数据集样本空间的特征维度；d 表示经过降维后的特征维度；k 为每个样本选取的近邻个数；$X_1, X_2, ..., X_n$ 表示 D 维的 n 个样本；$Y_1, Y_2, ..., Y_n$ 则表示降到 d 维后的 n 个样本；Y_i 为 X_i 在 d 维空间的嵌入表示；一般地，$k \le n-1$，$d \le D$。则具体算法共分为以下三步[51]：

Step1.局部邻域点搜索。使用欧式距离计算并选取距离样本点 x_i 最近的样本点作为样本点 x_i 的邻域点 $x_{i1},x_{i2},...,x_{ik}$，并假定 x_i 及其选择的邻域点构成的是线性超平面。

Step2.计算样本的局部重建权值矩阵 w。对于任意样本 x_i，用 step1 中选择的邻域点的线性组合重构表示 $\hat{x}_i$，即

$$\hat{X}_i = \sum_{j=1}^{n} \omega_{ij} X_j \tag{4-1}$$

系数 ω_{ij} 表示在重构时 x_i 的第 j 个邻域点所占的权重，当样本点 x_j 不是 x_i 的邻域点时，$\omega_{ij}=0$；并且满足 $\sum_j \omega_{ij}=1$。因此，要使得重构后的代价误差函数 $\varepsilon(W)$ 最小，即

$$\min \varepsilon(W) = \min \sum_{i=1}^{n} \left\| X_i - \sum_{j=1}^{n} \omega_{ij} X_j \right\|^2 \tag{4-2}$$

对于每一个样本点 x_i，其代价误差 $\varepsilon_i(W)$ 为

$$\varepsilon_i(W) = \left\| X_i - \sum_{j=1}^{n} \omega_{ij} X_j \right\|^2 = \left\| \sum_{j=1}^{n} \omega_{ij} (X_i - X_j) \right\|^2 = w_i^T G_i w_i \tag{4-3}$$

其中 $w_i = (\omega_{i1},\omega_{i2},...,\omega_{in})^T$；$G_i = (G_{jk})$，并且

$$G_{jk} = (X_i - X_j)^T (X_i - X_k) \quad (X_j，X_k是X_i的领域点) \tag{4-4}$$

因为 $\sum_j \omega_{ij}=1$，即 $1_n{}^T w_i = 1$。因此利用拉格朗日乘子法，最小化函数

$$f(w_i) = w_i^T G_i w_i - \mu(1_n{}^T w_i - 1) \tag{4-5}$$

由此求得

$$w_i = \frac{G_i^{-1} 1_n}{1_n{}^T G_i^{-1} 1_n} \tag{4-6}$$

其中，$1_n = (1,1,...,1)^T$ 得到局部重建权值矩阵 w。

Step3.通过局部重建权值矩阵 w 求解低维空间中的最佳映射。局部重建权值矩阵 w 反映了高维空间数据的局部拓扑结构；由于要保持数据的内在特性，因此 ω_{ij} 也用于低维空间中重建数据的嵌入坐标。即低维嵌入样本 y_i 的线性组合重构 $\hat{Y}_i$ 表示为

$$\hat{Y}_i = \sum_{j=1}^{n} \omega_{ij} Y_j \tag{4-7}$$

使用局部重建矩阵 W 计算嵌入 d 维空间的重建代价误差 $\varepsilon(Y)$，并使得 $\varepsilon(Y)$ 最小，即

$$\min \varepsilon(Y) = \min \sum_{i=1}^{n} \left\| Y_i - \sum_{j=1}^{n} \omega_{ij} Y_j \right\|^2 = \min \sum_{i=1}^{n} \left\| (I_n - W) Y_i \right\| = \min tr(Y^T M Y) \tag{4-8}$$

其中 $M = (I-W)^T (I-W)$，tr 表示矩阵的迹，I_n 为单位矩阵，且满足 $\frac{1}{n} YY^T = I_d$。由于矩阵的迹与特征值相关，因此该问题进一步转化为求取矩阵 M 最小非零特征值的问题[51]。设矩阵 M 按升序排列的 D 个非零特征值所对应的特征向量为 $u_1, u_2, ..., u_D$，则最终求得嵌入 d 维空间的样本数据集为 $Y = (u_1, u_2, ..., u_d)^T$。

LLE 算法把握数据的局部特征，通过对局部拓扑结构的描述从而在数据的高维空间中寻找最佳低维嵌入。该算法引入了从局部把握整体的哲学思想，其对于流形结构的优异降维效果，引起了学术界的重视和兴趣。LLE 算法具有如下的优点：

（1）LLE 算法更多的考虑了样本间的相互关系，因此降维后的数据样本能够保留原数据样本的内部特性。

（2）LLE 算法不需要反复迭代便可求出其解析的全局最优解，与其他非线性降维算法相比，时间复杂度小，效率更高。

（3）LLE 算法能够处理复杂的数据，在处理非线性结构数据时，降维效果明显优于线性降维算法。

（4）LLE 算法使用简单，在实际应用中输入参数简单，只需要配置邻域点个数 k 以及映射后的低维维度 d。

（5）LLE 算法使用时对数据的限制条件少，使用范围广泛。

当然，也由于 LLE 算法自身的特点也存在如下的缺点：

（1）邻域点数 k 的选择。邻域点数过大或过小都会影响降维的效果；因此需要研究者根据实际情况在实验中选择最佳领域点数。

（2）本征维度 d 的选择。本征维度 d 的选取决定使用几个独立参数来描述数据集的本质特征。现在缺乏一套完善的评估标准来选择 d。

（3）LLE 算法对噪声较为敏感。如数据采集设备的局限、自然条件等限制会造成数据存在噪声，而噪声的存在对于 LLE 算法有较大影响。

4.3 局部线性嵌入算法的改进

本小节将提出一种基于 K-Means 和均值限制的邻域点选择方式，该方法可以有效地克服局部线性嵌入算法中邻域点数 k 选择过大时造成的问题。

4.3.1 局部线性嵌入算法邻域点选择的问题

LLE 算法将数据的局部几何性质视为线性结构，将高维的流形结构数据划分为许多拥有线性结构的小平面块，然后以每个小平面为单位映射到低维空间中，再按照原数据集间的网路拓扑关系重新拼合起来，从而得到了高维结构数据的低维空间表示。因此原始数据通过 LLE 降维后可以保留其本质特征。如图 4-6 中：

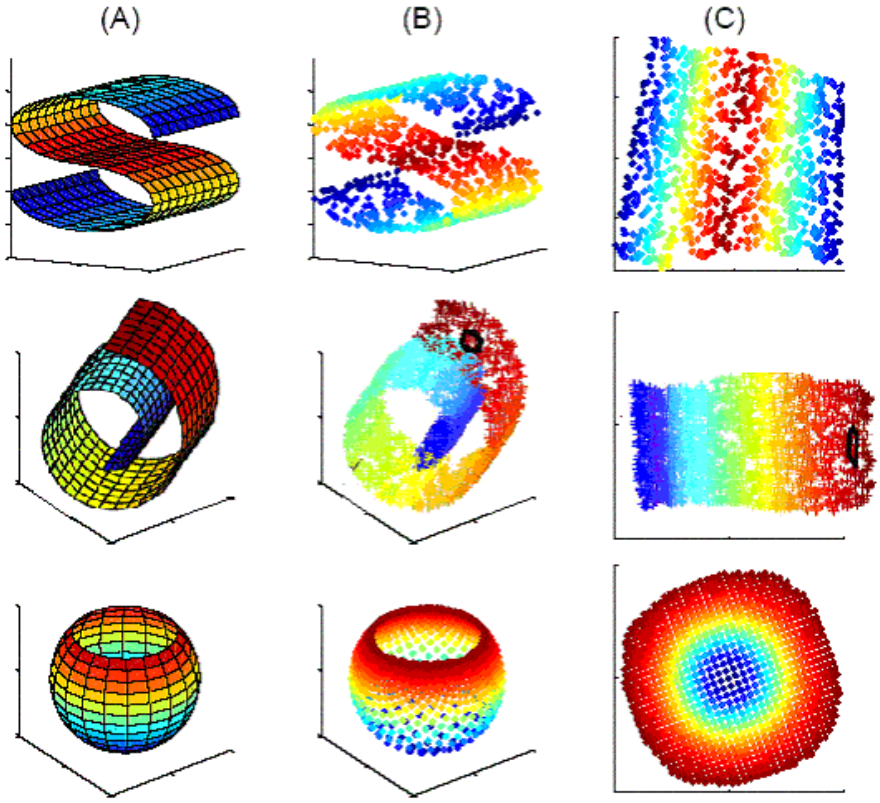


图 4-6 局部线性嵌入降维有效保持数据流形结构

图中将三维非线性数据映射到二维空间。图中第一二三列分别表示的是原始数据、采样数据和使用局部线性嵌入降维后的二维空间数据；这里，数据集的类别信息通过不同的颜色进行表示。从图中可以看出，把原始的三维数据集通过局部线性嵌入算法降维后，二维空间中的数据颜色分布依然较为明确集中即仍能保留相对独立的类别信息。黑色小圈如第二行中的图（B）所示，其映射到二维空间后仍分布在红色类别数据中，良好地保持原有数据的流形结构。因此，LLE 算法良好地保持了原流形结构中局部邻域间的互相关系，常用于分类挖掘中的降维预处理中。

LLE 算法是在保持高维数据样本点拓扑关系的基础上，将高维特征空间的数据映射到一个低维空间。因此，在维持拓扑关系的问题上，对每个数据样本的邻域点选择是相当重要的，它对最后的降维效果将产生显著的影响[52]。对于邻域点数的选择，如果邻域点数 k 选择太小，算法会将连续的流形结构割裂成若干不连通的子流形，产生“孔洞”现象。“孔洞”现象将原始流形结构的数据分成支离破碎的独立小区域，破坏了样本点在低维空间中的拓扑结构，从而导致映射将无法反映整体数据的内在特性。而若选取的邻域点数过大，则有可能会出现“短路边”现象，将原本距离较远的点作为自己的邻域点，无法体现样本的局部特性，从而破坏原数据流形的拓扑性质，使得降维结果表现得跟传统的 PCA 一样。如图 4-7 所示：

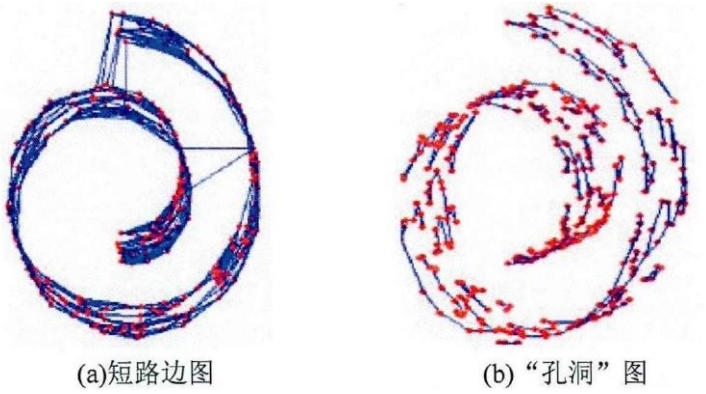


图 4-7 LLE 算法邻域点数选择不当引发的问题

如图(a)短路边图所示，当 k 值选择过大时，会选择距离样本点流形结构很远的点作为其近邻点，破坏了流形结构；而(b)“孔洞”图中，k 值过小，则会将原始数据集切分成很多不连通的小子块，破坏了其拓扑结构，从而导致低维嵌入时效果较差。因此如何选取最优邻域点数 k 是学者们对 LLE 算法研究中需要解决的一个重要问题。

为选取最优的邻域点数 k ，学者们展开了大量的研究。Kouropteva 和 Okun 等人首先提出了剩余方差[53]来度量“映射”的好坏。接着又提出了利用剩余方差来选取最优邻域点数的方法：直接法和分层法。国内的张兴福、吴学斌等同学也对邻域点数的选择做了相应的探讨和研究，并取得了一定的成果。本文将提出一种基于 K-Means 和均值限制的局部线性嵌入算法，可以有效克服因 k 选择过大而导致的“短路边”现象。

4.3.2 K-Means 聚类方法

K-Means 是聚类分析应用最为广泛的方法之一。聚类方法是将一个数据集按照某一设定的距离计算标准分割成不同的类或簇。对于这些形成的簇，尽可能使同一个簇内的数据对象距离较近且相似性高，而簇与簇之间的数据对象之间距离较远且差异性大。即使用聚类方法后相似数据对象紧凑地聚集在相同簇中，不相似数据对象尽量分离。聚类分析算法有多种，可根据数据类型、聚类目的和具体应用的不同进行相应选择[54]。

K-Means 算法的基本思想是初始随机给定 k 个簇中心，样本根据距离计算公式选择最近的簇中心并将其归入该簇中，完成第一次分配。之后，根据簇中样本计算各簇的质心并确定为新簇心。该算法的主要步骤如下：

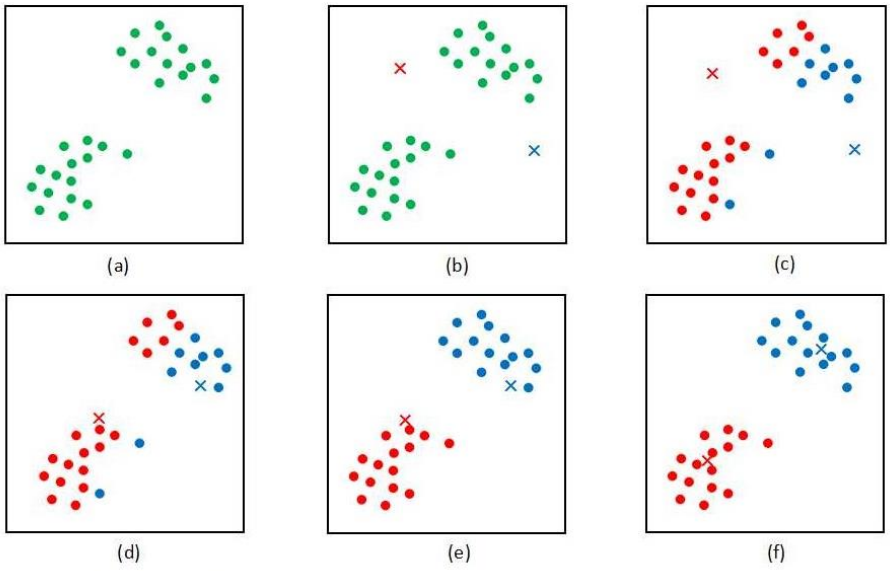
（1）为数据集初始化聚类簇中心。

（2）计算各样本到簇中心的距离，并将其分配到距离最近的簇中，作为该聚类簇的成员。

（3）计算各簇中成员的坐标均值，并将其作为新的簇中心进行下一步迭代。

（4）重复执行步骤（2）、（3）；知道各簇中心不再变换或聚类次数达到设定阈值为止。

下图 4-8 展示了对某个数据集进行 K-Means 聚类的过程：



(a) 未聚类的初始点集；

(b) 随机选取两个点作为聚类中心；

(c) 计算每个点到聚类中心的距离，并聚类到离该点最近的聚类中去；

(d) 计算每个聚类中所有点的坐标平均值，并将这个平均值作为新的簇中心；

(e) 重复(c),计算每个点到聚类中心的距离，并聚类到离该点最近的簇中去；

(f) 重复(d)，计算每个聚类中所有点的坐标平均值，并将这个平均值作为新的簇中心，直到满足要求。

图 4-8 K-Means 聚类实例图解

从上述的 K-Means 分析中可以看出，K-Means 将相似的样本点聚集到一个簇中，而这些相似的点也通常是样本点自身的邻域点，因此将 K-Means 算法运用于 LLE 算法的领域点选择的限制当中，可以有效地避免邻域点数 k 选择过大而造成的“短路边”现象。

4.3.3 基于 K-Means 和均值限制的邻域点选择

在基于 K-Means 和均值限制的邻域点选择中，均值表示为初始近邻点与样本点之间的距离均值。

这里以样本 x_i 为例，并假设选取的近邻点数 k=10，那么在传统局部线性嵌入算法中选取的近邻如下图 4-9 所示：

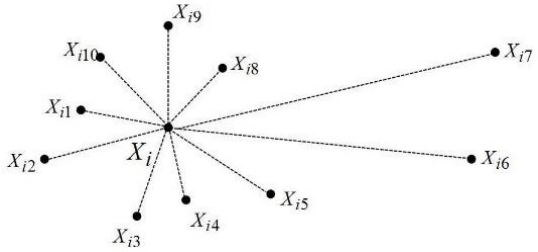


图 4-9 样本 X_i 选择 10 个近邻点

从图中可以看出，样本 x_i 选取的 10 个近邻距离参差不齐，差距较大；尤其对于近邻点 x_{i6} 和 x_{i7} 距离很大。这种距离较远、贡献较小的近邻被选取后将导致低维嵌入结果误差较大，最终影响降维的最终效果。

针对以上现象，在采用有均值限制的情况下，近邻的选择会得到较好的改善。如下图所示：

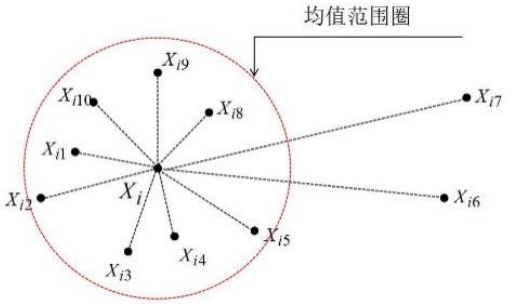


图 4-10 均值限制后，样本 X_i 剔除近邻点 x_{i6} 和 x_{i7}

通过均值的限制，有效剔除了距离较远的近邻点 x_{i6} 和 x_{i7}，剩下的邻域点与样本 x_i 较为相似且距离较近。通过自适应的选取邻域点有效避免了“短路边”现象的发生，具有较为良好的低维嵌入结果。

因此，综合之前的问题分析和方法介绍，基于 K-Means 和均值限制的邻域点选择的主要思想是：首先对样本数据集进行聚类分析；然后选择初始邻域点并计算其与样本点的距离均值；最后逐一判断初始邻域点与样本点距离是否小于均值或与样本点处于同一聚类簇中，若是则该点作为最终邻域点，否则不是。

为了方便叙述，定义如下符号：D 表示原始数据集样本空间的特征维度；d 表示经过降维后的特征维度；k 表示每个样本选取的近邻个数；$X_1, X_2, ..., X_n$ 表示 D 维特征的 n 个样本；$X_{i1}, X_{i2}, ..., X_{ik}$ 表示样本 X_i 的初始近邻点；D_{ik} 表示样本 X_i 与初始近邻点 X_{ik} 的距离；MD_i 表示样本 X_i 的初始近邻点平均距离；$Y_1, Y_2, ..., Y_n$ 则表示降到 d 维后的 n 个样本；Y_i 为 X_i 在 d 维空间的嵌入表示；一般地，$k \le n-1$，$d \le D$。则基于 K-Means 和均值限制的 LLE 改进算法如下：

（1）对数据集进行聚类。使用基于欧式距离的 K-Means 算法对样本数据集进行聚类，使得相互邻近的相似样本点处于相同聚类簇中。

（2）选取每个样本点的初始近邻点。根据设定的领域点数 k 选取距离自己样本最近的 k 个样本作为自己的初始邻域点。即样本 X_i 的初始近邻为 $X_{i1}, X_{i2}, ..., X_{ik}$。

（3）计算初始近邻的平均距离。对于样本 X_i 的初始近邻为 $X_{i1}, X_{i2}, ..., X_{ik}$，计算近邻点距离样本 X_i 的平均距离 MD_i，即

$$MD_i = \frac{1}{k}(D_{i1} + D_{i2} + ... + D_{ik}) \tag{4-9}$$

（4）选取每个样本的最终邻域点。对于选取样本 X_i 的最终邻域点，需满足一下两个条件之一：若初始近邻点 X_{ik} 同样本 X_i 的距离 D_{ik} 小于平均距离 MD_i，或者样本 X_i 与初始近邻点 X_{ik} 处于同一聚类簇中，则 X_{ik} 为 X_i 的最终邻域点；否则 X_{ik} 不是 X_i 的最终邻域点。

（5）计算样本的局部重建权值矩阵 W。这里跟上小节中局部线性嵌入算法的步骤 2 相同，即：将样本 X_i，用它的邻域点的线性组合重构 $\hat{X}_i$，并通过最小化重构后的代价误差求取局部重建权值矩阵 W。

（6）通过局部重建权值矩阵 W 求解嵌入低维空间中的最佳映射。将低维嵌入样本 Y_i 通过它的最终邻域点线性组合重构为 $\hat{Y}_i$，并使用局部重建矩阵 W 计算嵌入 d 维空间的重建代价误差 $\varepsilon(Y)$，并使得 $\varepsilon(Y)$ 最小，最终求取数据样本集 $X_1, X_2, ..., X_n$ 嵌入到 d 维后的样本集 $Y_1, Y_2, ..., Y_n$。

总得来说，改进后的局部线性嵌入算法主要对样本的邻域点选择做了改进与优化，而计算局部重建权值矩阵和最终的求解嵌入低维空间的最佳映射步骤与传统算法并无二异。图 4-11 是改进后的局部线性嵌入算法的流程图：

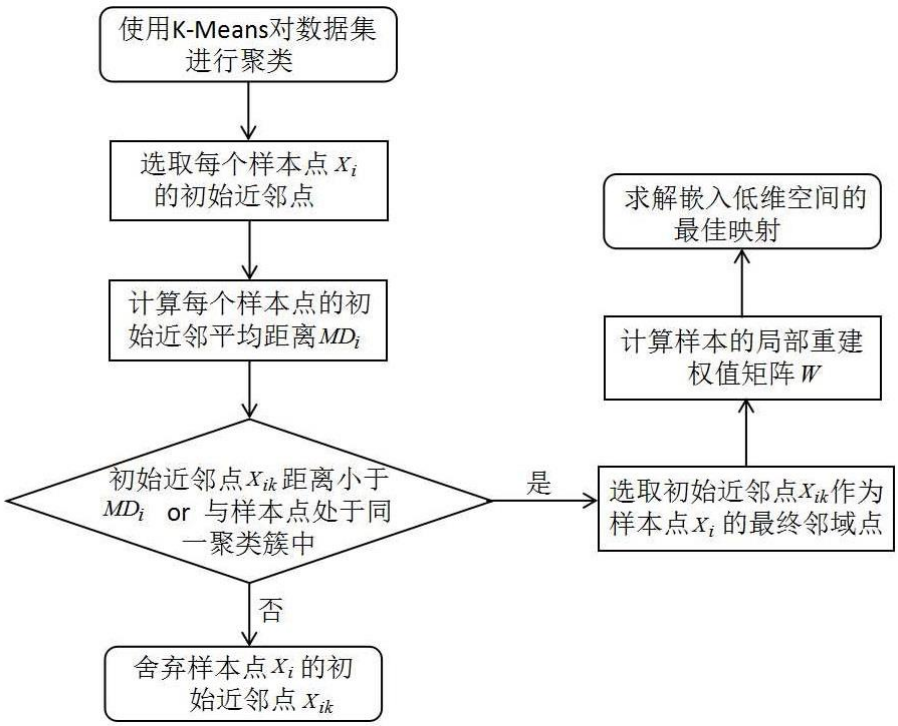


图 4-11 改进后的局部线性嵌入算法流程图

流程图中可以看出，通过增加聚类和样本点初始近邻距离均值的限制，从而舍弃掉一部分近邻点。通过这两者的限制可以有效的避免样本点 X_i 选取到距离较远的样本作为邻域点。通过自适应的为每个样本点选择合适的最终邻域点，可以更好地保持样本的局部几何结构。

若设输入数据集为 data, 样本选择近邻个数参考为 neighbors，最终的降维维度为 components, 聚类簇个数为 clusters, 则实现改进的 LLE 降维算法的伪代码如下所示：

```
Algorithm improved_LLE(data,neighbors,components,clusters){
    //使用 KMeansInfo 函数对数据集进行聚类分析
    c = KMeansInfo(data,clusters)
    for(int i=0;i<=data.length;i++)
    {
        //使用 getInitialNeighbor 函数选取样本点的初始近邻样本
        initialNeighbors = getInitialNeighbors(data[i],neighbors);
        //根据初始近邻样本计算平均距离
        mean = getNeighborsMean(initialNeighbors);
        //根据 K-Means 和均值的限制选择获取样本点的最终近邻样本
        realNeighbors[i]= getRealNeighbors(initialNeighbors,mean,c[i]);
```

```
    }
    //根据 RecWeightsMatrix 函数计算数据集的局部重建权值矩阵
    RWMatrix = RecWeightsMatrix(data,realNeighbors[i]);
    //根据用户输入的 components 以及重建权值矩阵计算最终的低维映射
    result = GetLowDimens(data,RWMatrix,components);
    //返回最终结果 result
    return result;
}
```

如图 4-12 是对改进后的局部线性嵌入的演示：

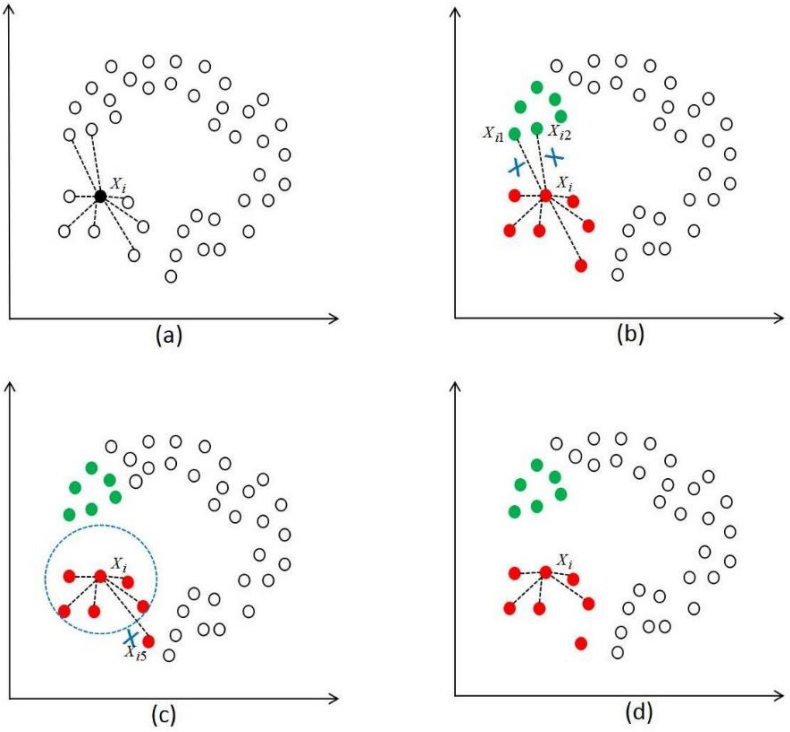


图 4-12 改进的局部线性嵌入算法实例图解

当 k =8 时，在图(a)中，样本 x_i 选取了距离最近的 8 个样本点作为初始近邻点；图(b)中，红色与绿色属于两个不同的聚类簇，因为样本点 x_i 与两个近邻点属于不同的聚类簇，因此剔除属于绿色类的近邻点 x_{i1} 、 x_{i2} ；而在图(c)中，因为近邻点距离样本 x_i 的平均距离为蓝色圆圈所示，因此剔除近邻点 x_{i3} ；图(d)显示了样本 x_i 的最终邻域点。从图(d)可以显示出，样本 x_i 的最终邻域点与样本 x_i 相近且相似，保持了局部拓扑结构，最终有利于嵌入低维空间后保持流形的内部特性，有效阻止了“短路边”现象的发生。

4.4 实验结果

4.4.1 脑电波时频域信号的转换

实验中的EEG脑电波数据是在采样频率5000HZ下的一段时间的脑电波信号。信号由 8 名癫痫病患者发作期间数据和发作间歇期数据两部分组成，存储大小为40GB左右。主要目的是通过对这两类数据进行训练，从而建立分类模型预测病人是否为癫痫病发作状态。这八名患者的数据存储大小如下表所示：

表 4-1 数据集存储量统计

患者 1	患者 2	患者 3	患者 4	患者 5	患者 6	患者 7	患者 8
1.35GB	3.78GB	4.26GB	1.62GB	13.12GB	6.70GB	9.46GB	2.24GB

图 4-13 展示了一名患者在一个周期内的脑电波时域信号波形图：

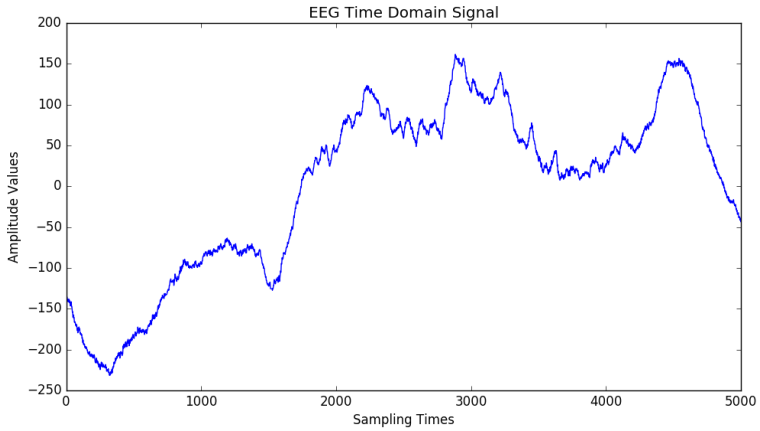


图 4-13 患者的时域信号波形图

从图中可以看出，脑电波的时域信号仅仅反应了信号在时间域上的状态，并不能直观的反映脑电波信号在各个节律的分布情况。由于时域分析的局限性，因此实验中使用快速傅里叶变换，将脑电波时域信号转换到频域再进一步的分析。相对应的频域信号则如图 4-14 所示：

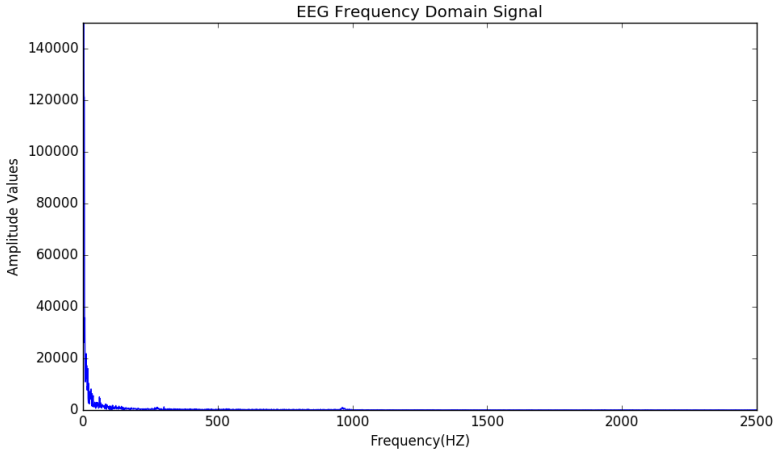


图 4-14 患者的频域信号波形图

从图中可以看出脑电功率随频率的变化，而频域信号 500HZ 以上几乎为 0，可视其为噪音。而对脑电波频域信号 0~150HZ 的情况如图 4-15 所示：

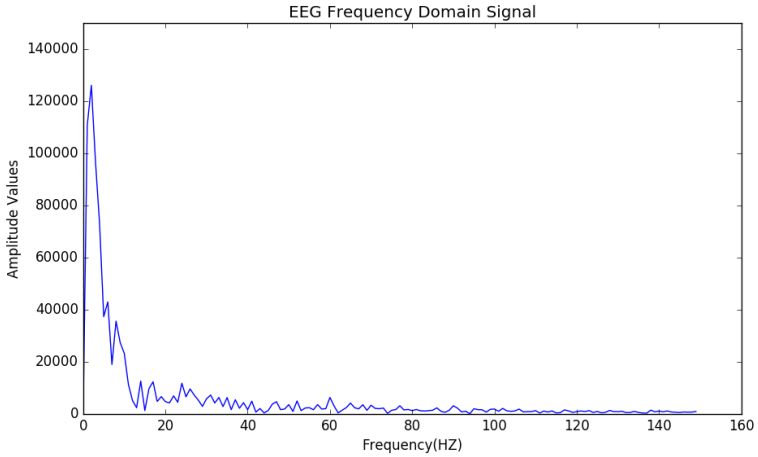


图 4-15 患者 0-150HZ 的频域信号波形图

可以看到，随着频率的增加幅度在骤减并接近于 0。通过进一步的实验分析表明：当对 0~100HZ 的频域数据承载的脑电波信号最多，对其进行分析时效果最佳。因此接下来的实验是对数据的 0~100HZ 脑电波频域信号进行分析。

4.4.2 性能结果对比与分析

实验中，采用 SVM 对脑电波的频域信号进行预测分类；在使用 SVM 分类之前，对频域信号进行四种不同处理：不处理、PCA 降维处理、LLE 降维处理、改进的 LLE 降维处理。然后对这四种不同处理方法的最优结果进行各方面性能指标对比。

（1）时间对比

表 4-2 各方法的挖掘时间对比

降维 / 患者	无	PCA 降维	LLE 降维	改进 LLE 降维
患者 1	44min 17s	10min 49s	12min 22s	14min 41s
患者 2	80min 24s	20min 12s	23min 40s	25min 01s
患者 3	117min 54s	28min46s	30min 33s	33min 45s
患者 4	64min 08s	16min 33s	20min 44s	21min 57s
患者 5	376min 22s	74min 43s	75min 20s	77min 29s
患者 6	172min 35s	29min 02s	31min 56s	34min 12s
患者 7	123min 20s	32min 08s	34min 42s	37min 54s

表 4-2 各方法的挖掘时间对比

患者 8	75min 37s	13min 09s	16min 45s	18min 02s
总计	1054min 37s	225min 22s	247min 02s	264min 01s

从表中可以看出，当不使用降维时的时间最长的，使用 PCA 降维时的时间是最短的，几乎是不降维处理时使用的五分之一时间。原因是使用 SVM 进行分类预测时，若属性维度过高会大大的增加计算复杂度，因此对频域信号进行降维是非常有必要的。而使用改进的 LLE 算法比传统 LLE 算法时间稍微长一些，原因是要对数据集进行聚类和均值计算，但时间长度增加并不明显。

（2）AUC 面积对比

表 4-3 各方法 AUC 面积结果对比

降维 / 患者	无	PCA 降维	LLE 降维	改进 LLE 降维
患者 1	0.886	0.833	0.852	0.873
患者 2	0.912	0.857	0.870	0.892
患者 3	0.898	0.845	0.859	0.874
患者 4	0.894	0.816	0.841	0.870
患者 5	0.933	0.865	0.887	0.906
患者 6	0.925	0.857	0.872	0.894
患者 7	0.897	0.823	0.844	0.869
患者 8	0.918	0.842	0.853	0.881
平均	0.908	0.842	0.865	0.882

从表中可以看出，对频域信号不使用降维处理而直接进行 SVM 分类时，AUC 面积达到了 0.933。而使用线性降维方法 PCA 对其进行降维后 AUC 面积下降了接近 6 个百分点。从对比可以看出使用 LLE 降维算法降维时，AUC 面积比 PCA 降维算法平均高出 2 个多百分点，这说明对于脑电波频域信号非线性降维方法比线性降维更好。而改进后的 LLE 降维算法又比传统 LLE 降维算法的 AUC 面积平均提高了接近 2 个百分点，这表明改进后的 LLE 降维算法的邻域点选择更加合适，低维嵌入数据更好地保持了高维空间的流形结构。

由上图两表可以发现，使用 PCA 降维比 LLE 降维所花费的时间更短，这是因为线性降维实现更为简单，计算复杂度较低。而使用 LLE 降维对频域信号进行降维比使用 PCA 降维时 AUC 面积更大，平均提升 2 个百分点；在此基础上，改进

的 LLE 算法比传统 LLE 算法 AUC 面积平均又提升了 2 个百分点；因此改进的 LLE 算法虽然在计算时间上比传统 LLE 算法稍长一些，但对于降维效果来说，改进的 LLE 算法更为优异。

（3）损失率。这里的损失率是指嵌入到低维空间的重建代价误差与真实值的差距。

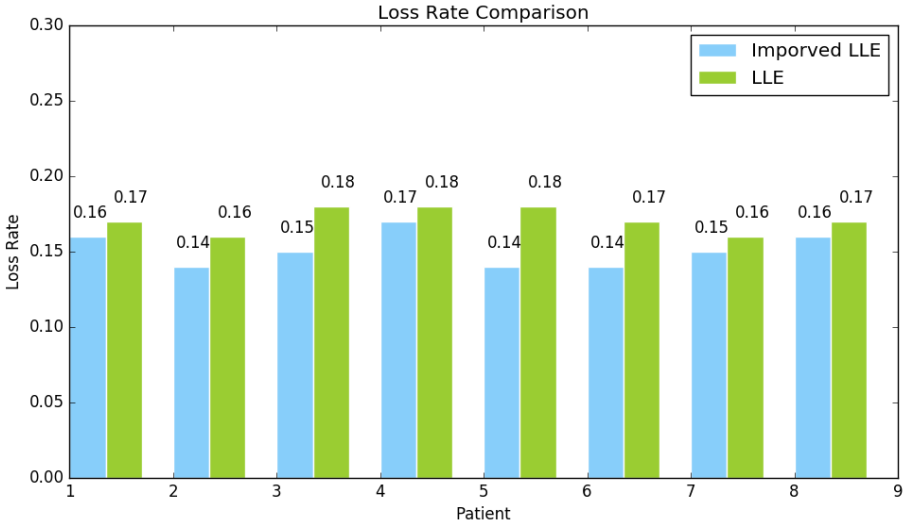


图 4-16 降维损失率对比图

从图中可以看到，改进的 LLE 算法嵌入低维空间的重建代价误差都在一定程度上小于未改进的 LLE 算法。在低维空间用邻域点线性表示样本点时的重建代价误差更低即损失率降低，说明高维空间中样本选择的邻域点更好反应了数据的局部拓扑结构，在嵌入低维时更好地保持了数据的流形结构与内部特性，因此改进的 LLE 算法的降维效果更佳。

4.5 本章小结

本章主要研究了 EEG 脑电波频域信号的降维处理方法。先介绍了几种常用的降维方法，包括线性降维方法和非线性降维方法。由于频域信号是复杂的非线性数据，因此选用了局部线性嵌入降维算法。之后对局部线性嵌入降维算法进行了深入的研究。在局部线性嵌入降维算法中邻域点的选择很重要，不适当的选择会产生“短路边”现象和“孔洞”现象。因此本章第三小节对邻域点的选择提出了改进，即基于 K-Means 和均值限制的邻域点自适应选择方法。第四小节针对脑电波数据进行实验并展开对比分析。

第五章 健康大数据预处理系统的设计与实现

本文之前的内容主要介绍了数据挖掘预处理经常使用到的理论技术并对部分方法进行了创新性的改进；本章节则侧重于将预处理技术应用于实际的数据挖掘中，着重介绍实际应用中预处理分析的架构设计以及核心实现。

5.1 需求分析

本文主要将数据挖掘中的预处理技术应用于数据挖掘实际应用中，从而提高下一步数据挖掘的准确率及效率。接下来针对两个数据集即“人口死亡”数据集和”和“癫痫病脑电波”数据集进行需求分析。

（1）人口死亡数据集

这里对国家卫生局收录的“人口死亡”数据集进行大数据处理分析，其目的在于：将数据挖掘技术应用于实际生活产生的健康数据中，通过对“自杀”和“非自杀”两类人群的相关详细信息的研究与分析，探讨自杀人群的明显特点，即寻找哪类人群容易产生自杀的倾向，从而指导国家相关部门采取相应的措施减少社会中自杀情况的发生。

在对人口死亡数据集进行数据挖掘前进行合理有效地预处理是有必要的。首先，收集的数据来自于多个数据库，因此需要将多个数据库整合到一起。其次，数据集信息量巨大，但并不是所有的数据都会对之后的数据挖掘有用，因此需要进行数据特征选择。再次，选择后的属性可能存在冗余等问题，因此需要再次对属性进行分析并筛选。最后，还需对数据集中的缺失值进行适当地处理、规范数据集使其适合进一步的挖掘分析。

（2）癫痫病脑电波数据集

这里将对癫痫病脑电波数据集进行大数据处理分析，其目的在于：将大数据处理技术应用于医学研究，通过挖掘癫痫病患者在发作期与间歇期的脑电波的不同，从而建立预测模型预测病人是否为癫痫病发作状态。

癫痫病患者的 EEG 脑电波数据是在采样频率 5000HZ 下的一段时间的脑电波信号，这就意味着每位癫痫病患者每秒钟采样的脑电波数据高达 5000 维。若直接对这些庞大又复杂的数据进行挖掘分析将会极大地降低处理效率。除此之外，数据采集时产生的不可避免的噪音会影响数据挖掘结果的可靠性，因此寻找合适的预处理技术对脑电波数据集进行预处理是十分有必要的。在数据预处理分析中不仅需剔除数据集中的噪音数据，还需进一步有效提高大数据处理分析的效率。

5.2 总体设计

在健康大数据处理系统中，主要分为“人口死亡处理”和“癫痫病脑电波处理”两个核心模块。如图 5-1 所示：

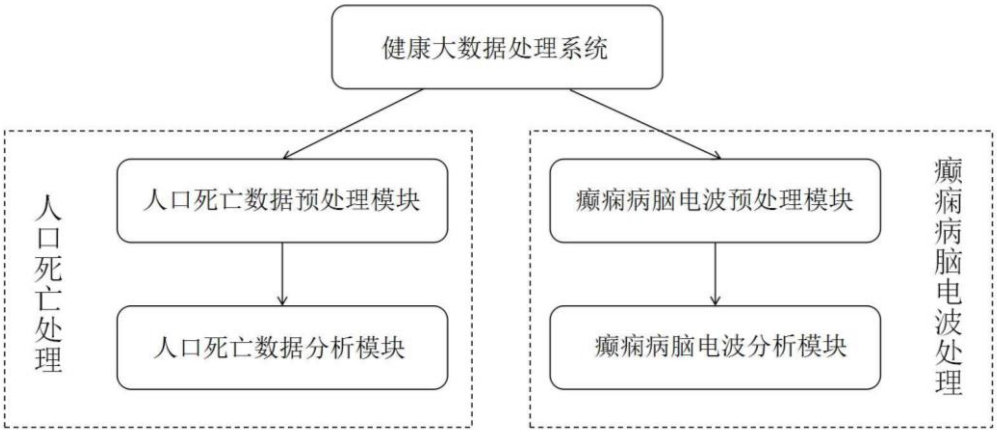


图 5-1 系统的总体框架设计

在健康数据的挖掘过程中，不同的健康数据集其采用的预处理方法和数据挖掘方法都会有所不同。因此对于每一个大的模块，又将其分为数据预处理模块和挖掘分析模块。每一个数据预处理模块又包含若干预处理步骤。在数据挖掘中，需要根据数据集的收集方式、信息特点、挖掘需求等因素综合考虑制定预处理使用的技术和方法。

因此，针对数据集的独有特点，人口死亡数据预处理模块的功能设计如图 5-2 所示：

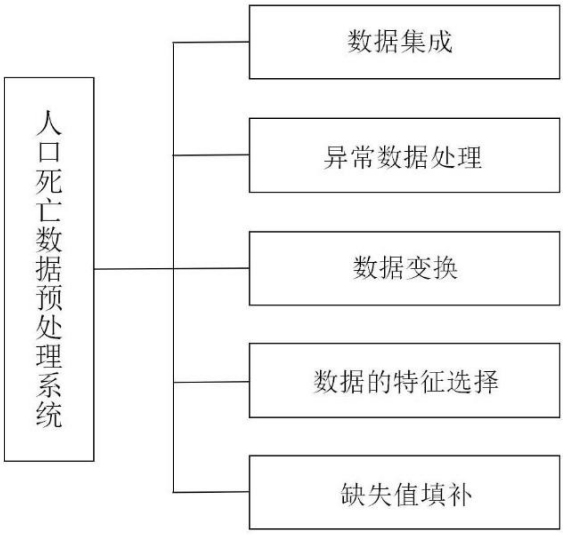


图 5-2 人口死亡数据预处理模块的功能设计

在此模块中，主要对人口死亡数据集进行五个步骤的预处理，依次为数据集成、异常数据处理、数据的特征选择、数据变换以及最后的缺失值填补。

由于收集的数据来自多个不同的数据库，因此若直接将多个数据库的数据直接整合在一起会出现错误或者冗余；这是由于每个数据库的属性元数据以及元数据对应的尺寸范围也不相同造成的，因此数据集成是对数据进行分析的第一个重要步骤。其次，由于人为或系统等原因导致部分数据为不合逻辑或明显错误的异常数据。因此，对数据集进行相关的异常数据检测和处理相当有必要的。之后在处理时还需对数据集进行适当的变换，如对连续的属性进行离散化处理、把属性数据按一定的比例缩放等。然后，再对数据集进行特征选择，通过分析各个属性特征和类别的相关性权重，选择合适的特征。最后，通过现代的缺失值填补技术对数据集的缺失值进行填补，为之后的数据挖掘分析做准备。

而在对癫痫病脑电波数据集进行数据挖掘时，其模块的功能设计如图 5-3 所示：

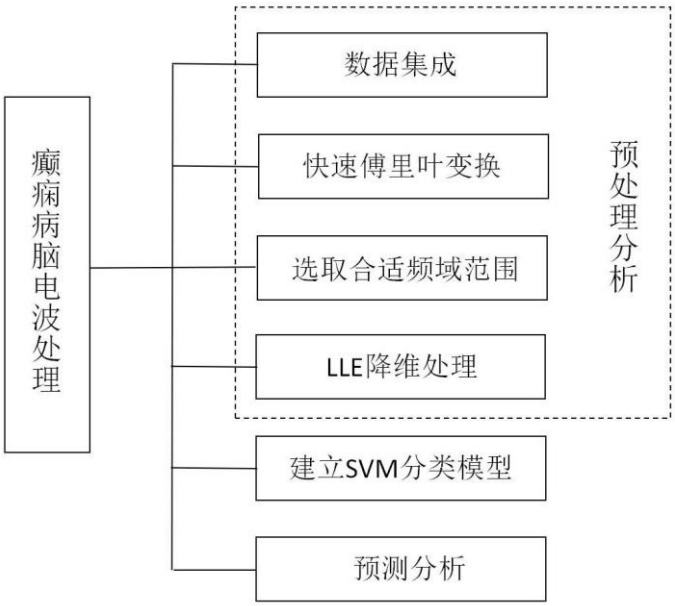


图 5-3 癫痫病脑电波处理模块的功能设计

在对癫痫病脑电波进行预处理分析时使用到了预处理技术中的数据集成、数据清理、数据归约所包含的技术方法。如图 5-3 所示，首先从文件中读入 EEG 脑电波数据，统一数据源元数据的属性，并将读入的数据以合理的方式存放在多维数组中，便于之后的数据处理。这里存放的数据是时域上的脑电波数据，而时域上的数据不便于后期的观察，因此将数据通过快速傅里叶变换转换成频域上的信号，并分析各频段脑电波信号的分量；总的来说，频域分析相比时域分析更为的直观有效。当然，得到的频域信号的高频部分则是噪音信息，通过选取合适的频域信

号范围可以有效的剔除脑电波中的噪音。之后，再在合适的频域信号上使用 LLE 进行降维处理，从繁杂的数据中挖掘出有用的信息，降低数据处理的维度，并且从一定程度上提高挖掘效率。最后，建立 SVM 分类预测模型对数据进行分类预测。

以上的每一个步骤都不是互不相关的，每一步处理的微小差异都将对下一步的处理造成一定的影响。比如在“选择合适频域范围”这一步中，频域范围的选择将会对下一步的“LLE 降维处理”的结果造成影响；同时，频域范围选择的变换也会使得 LLE 降维中的最优参数发生变化。而“LLE 降维处理”的参数的选择变化也会对之后的 SVM 分类结果造成影响。因此，对 EEG 脑电波数据的处理分析也是一个不断选择、不断调参的循环过程，直到得到的预测结果满足限定的要求为止。

5.3 详细设计与实现

本章的前两小节分别对健康大数据预处理系统的需求分析和总体框架设计进行了讲解，这一小节将对“人口死亡”数据集预处理模块和“癫痫病脑电波”数据集预处理模块的具体实现进行详细说明。

5.3.1 人口死亡数据集预处理模块

1. 数据集成

由于数据集来自 4 个不同的数据库，由于数据库的设计与元数据的属性单位规定等差异导致数据集存在不一致问题。如果直接将多个数据集整合到一起必然会导致数据的错乱，无法进行下一步分析。因此，数据集成将来自多个不同数据库存储的数据——合并，并将其存放在一个一致的数据存储中。表 5-1 显示了 4 个数据库的部分不一致问题。

表 5-1 四个数据库的部分不一致问题

数据库 / 表示属性	数据库 1	数据库 2	数据库 3	数据库 4
性别	0(男)/1(女)	male/female	male/female	male/female
户口类别	country/urban	0(城镇)/1(农村)	country/urban	country/urban
身高	单位(cm)	单位(m)	单位(cm)	单位(m)
残疾	0(正常)/1(残疾)	0(正常)/1(残疾)	0(正常)/1(残疾)	yes/no
身份证命名	ID_num	ID_num	ID_card	ID_num

续表 5-1 四个数据库的部分不一致问题

死亡地点	home/outside	home/outside	1(家)/0(外面)	1(家)/0(外面)
学历	elemente(小学) junior(初中) high(高中) bachelor(大学) master(硕士) doctor(博士)	0(小学) 1(初中) 2(高中) 3(大学) 4(硕士) 5(博士)	0(小学) 1(初中) 2(高中) 3(大学) 4(硕士) 5(博士)	elemente(小学) junior(初中) high(高中) bachelor(大学) master(硕士) doctor(博士)

从表中可以看到：数据表 1 中，对于“性别”这一属性，男性用 0 表示，女性用 1 表示；而在其他数据表中，男性使用英文 male 表示，女性则使用 female 表示；而对于“户口类别”这一属性，数据表 1、数据表 3 和数据表 4 均使用 urban 作为城镇户口，country 作为农村户口；而数据表 2 则使用 1 表示农村户口，0 表示城镇户口。除此之外，对于“身高”这一属性，4 个数据库的取值单位不统一；对身份证这一属性命名也不相同，有的数据库使用 ID_num，而有的则使用 ID_card。针对以上问题表明在对多个数据库进行集成时，需对数据库进行不一致问题处理。

通过对 4 个数据库的合并以及统一规范存储，集成的数据表基本信息见表 5-2 所示：

表 5-2 数据库的信息统一

序号	字段	含义	类型
1	id	主键	bigint(8)
2	ID_card	身份证号码	int(18)
3	name	姓名	varchar(20)
4	age	年龄	int(4)
5	sex	性别	boolean
6	nativePlace	户口类别	boolean
7	disability	是否残疾	boolean
8	education	学历程度	int(4)
9	occupation	从事行业	varchar(30)
10	placeOfDeath	死亡地点	varchar(40)
11	address	居住地点	varchar(50)
12	yearOfDeath	死亡年份	int(4)
13	monthOfDeath	死亡月份	int(4)
14	weekOfDeath	死亡星期	int(4)
15	sexualOrientation	性取向是否正常	boolean

续表 5-2 数据库的信息统一

16	bornTime	出生时间	smalldatetime
17	maritalStatus	是否结婚	boolean
18	activityLove	是否喜欢运动	boolean
19	congenitalDisease	是否有家族遗传史	boolean
20	injuryInWork	是否在工作中受过伤	boolean
21	infant	是否是婴儿	boolean
22	mentalDisease	是否有心理疾病	boolean
23	mannerOfDeath	死亡方式	varchar(30)

2.异常数据处理

在数据的收录过程当中，许多信息存在着不符合实际情况的错误。因此这里的异常数据处理，可以将这些错误数据进行清理，消除噪声、光滑数据集。

首先清理的是数据集中的年龄字段。由于人类的正常年龄是 0~120 岁，因此将数据集的年龄进行可视化分析如图 5-4 所示：

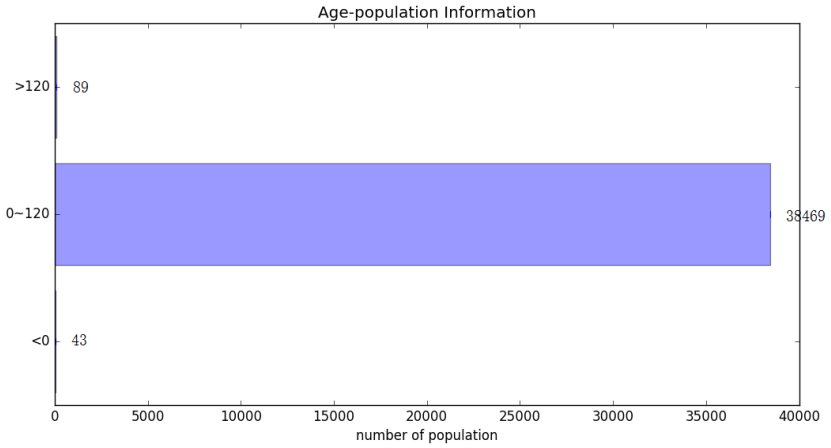


图 5-4 数据集年龄分布统计图

图中可以看到，在人类正常年龄范围 0~120 岁之外，依然存在着少量的数据，因此需对这些噪音数据进行删除清理。

而在对出生年份进行异常数据检测时，通过出生年份的可视化图 5-5：

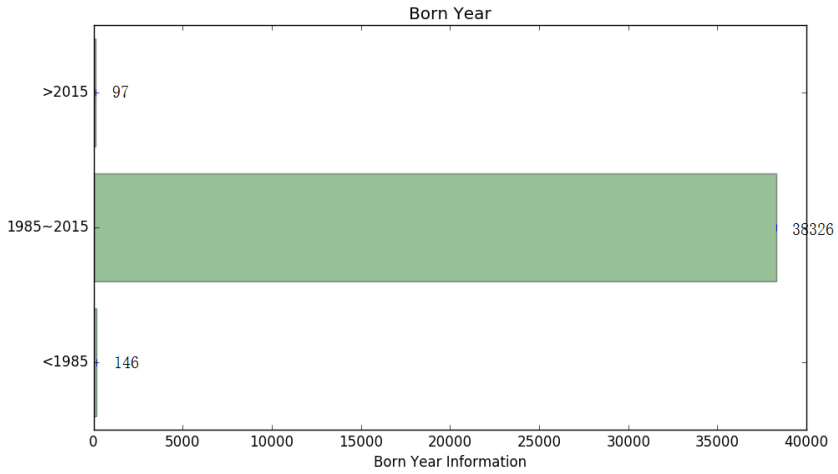


图 5-5 数据集出生年份分布统计图

如图可以看出，在 1895 年之前出生的异常数据依然存在 146 条，而在 2015 年之后出生的异常数据依然有 97 条，因此对这两类异常数据进行删除清理，从而达到光滑噪音的效果。

2. 数据变换与规范化

由于数据挖掘对数据的格式有特定的限制，因此这一步骤是对数据集的格式进行合适的转换，使数据集能够进行下一步的处理。

在数据变换中，首先对数据集的死亡地点（placeOfDeath）和居住地点（address）这两个字段进行概念分层。即将具体的地点用大范围的地区号来表示。这样处理使得繁杂无规律的地点能够清晰有条理的用大区域来表示，便于之后的统计及挖掘分析。划分方式如下表 5-3 所示：

表 5-3 地区划分详情

地区编号	地区	涵盖省份	统计数量
1	华东地区	山东、江苏、安徽、浙江、福建、上海	6330
2	华南地区	广东、广西、海南	4998
3	华中地区	湖北、湖南、河南、江西	5214
4	华北地区	北京、天津、河北、山西、内蒙古	6397
5	西北地区	宁夏、新疆、青海、陕西、甘肃	5762
6	西南地区	四川、云南、贵州、西藏、重庆	5985
7	东北地区	辽宁、吉林、黑龙江	4640

其次是对连续的“年龄”属性的离散化处理。连续属性离散化的实质是在信息丢失最小化的限制下，将属性值的连续区间转换成少数有限值，从而有效提高分类算法计算效率。离散化的处理方式如图 5-6 所示：

图 5-6 数据集的年龄离散化处理

将年龄段分为童年、少年、青年、中年、老年并分别用 1、2、3、4、5 表示。

最后的转换是处理数据集中的“从事行业”以及“死亡方式”属性。从表 5-2 可以看出，这两个字段是用字符串表示。因此需要把它们转换成适合数据挖掘的方式。转换的规则如下表 5-4 所示：

表 5-4 地区划分详情

属性	转换方式
occupation	1现实型 / 2研究型 / 3艺术型 / 4社会型 / 5企业型 / 6常规型
mannerOfDeath	1自杀 / 0非自杀

因为接下来要对自杀和非自杀这两类数据进行挖掘分析，因此这里把正常死亡、他杀、意外死亡等非自杀方式都归入非自杀范围。

3.数据的特征选择

由于接下来会对“死亡方式”属性中的自杀和非自杀进行缺失值填补，因此这里对影响该结果的特征进行选择。这里采用 Relief 算法对各个特征相关性的权重进行计算，并将权重小于某个阈值的特征移除。

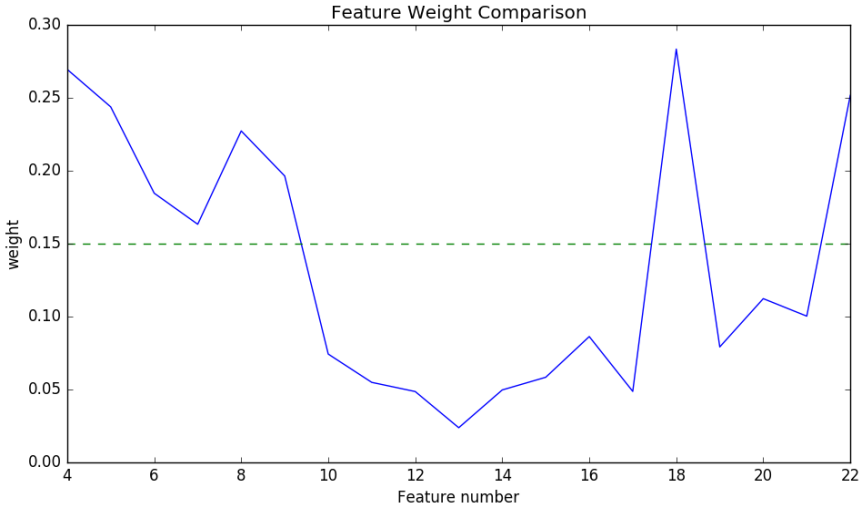


图 5-7 Relief 算法对特征进行权重分析

如图 5-7 所示为 Relief 算法计算的特征属性权重比较，这里选取权重大于 0.15 的属性作为下一步缺失值填补的特征属性；即年龄、性别、户口类别、是否残疾、学历程度、从事行业、是否喜欢运动、是否有心理疾病这 8 个特征。

4.缺失值填补

在数据集中的“死亡方式”这一属性字段存在着缺失值现象，因此需通过合适的方式对缺失值进行填补。这里使用现代的缺失值填补方法，简而言之，便是建立预测模型对缺失值进行预测分析并填补。在本文的第三章对现代的缺失值填补方法进行了详细的讲解；并通过对比确立了使用随机森林对数据缺失值进行填补。由于随机森林方法受到非平衡数据集的限制，因此采用 SMOTE 算法对数据集进行平衡性优化。本文的第三章第三小节对 SMOTE 算法进行了合理改进，这里便将改进后的 SMOTE 算法应用于人口死亡数据集中。

（1）非平衡数据集改善

如表 5-5 为使用改进的 SMOTE 算法平衡数据集的函数示例。

表 5-5 改进 SMOTE 算法的函数示例

Function	Description
getNegativeData	获取少数类样本
getBanRate	计算非平衡率
getCenterPoint	获取少数类样本的重心
getRandomSamples	随机获取两个少数类样本
createNewSample	合成新样本
calcuBalance	计算非平衡率
fillSet	填充少数类样本

若设输入数据集为 data, 非平衡率为 balanceRate，则实现改进的 SMOTE 算法的伪代码如下所示：

```
Algorithm improved_Smote(data,balanceRate){
    //初始化合成样本数据集
    newData = []
    //使用 getNegativeData 函数获取数据集的少数类样本
    negData = getNegativeData(data)
    //获取少数类样本的重心
    c = getCenterPoint(negData)
    //通过 getBanRate 计算数据集的非平衡率
```

```
    rate = getBanRate(data,negData)
    //当非平衡率小于设定阈值时，合成新的样本
    while(rate > balanceRate)
    {
        //随机获取两个少数类样本
        randData = getRandomSamples(negData)
        //使用 createNewSample 函数合成新样本
        newSample = createNewSample(randData,c)
        //将合成样本放入到合成样本数据集中
        newData.append(newSample)
        //计算现在的非平衡率
        rate = calcuBalance(data,negData,newData)
    }
    //根据 fillSet 将原始数据集和合成的新数据集合并
    result   = fillSet(data,newData)
    //返回最终数据集
    return result;
}
```

如上所示，在非平衡率 balanceRate 为 0.4 时，使用 improved_Smote 函数合成新的少数类样本并添加到原始数据集中。接下来便可以使用随机森林方法对缺失值进行建模分析并填补缺失值。

（2）随机森林填补缺失值

在使用随机森林建立分类模型时需要不断的调整参数，使得其分类预测效果达到最佳。表 5-6 为 cv=10 时，随机森利各参数输入的平均性能对比：

表 5-6 随机森林不同参数的性能对比

参数	分类精度	AUC 面积	F1-Measure
n_est=600,max_fea=5	0.8349	0.8492	0.8329
n_est=600,max_fea=6	0.8171	0.8346	0.8515
n_est=800,max_fea=5	0.8346	0.8607	0.8483
n_est=800,max_fea=7	0.8674	0.8312	0.8628
n_est=1000,max_fea=6	0.8843	0.8776	0.8652
n_est=1000,max_fea=7	0.8472	0.8548	0.8045

续表 5-6 随机森林不同参数的性能对比

n_est=1000,max_fea=8	0.8632	0.8231	0.8617
n_est=1500,max_fea=6	0.8753	0.8247	0.8715
n_est=1800,max_fea=5	0.8444	0.8767	0.8201
n_est=2000,max_fea=6	0.8417	0.8085	0.8117

表中 n_est 表示决策树棵树，而 max_fea 则表示分类时选择的最大特征数。从表中可以看出当 n_est=1000，max_fea=6 时的分类精度和 AUC 面积最佳，F1-Measure 位于第二且仅次于第一甚少。综合考虑，采用这组参数作为随机森林的输入参数进行分类预测建模从而对“死亡方式”属性的缺失值进行填补。

5.3.2 癫痫病脑电波数据集预处理模块

1.数据导入与集成

在实验中，癫痫病患者的 EEG 脑电波数据都存放在 MAT 二进制文件中，如图 5-8 所示：

图 5-8 癫痫病脑电波数据集

图中所示为 Patient1 的脑电波数据文件夹，从命名可以看出，ictal 表示癫痫病发作期间的脑电波数据，而 interictal 则表示间歇期间的脑电波数据，通过 python 中的数值计算库 scipy 中的 loadmat 读取 MAT 文件并保存在数组中；其中每个 MAT 文件作为一行代表一个维数为 5000 列的采样数据，因此每位病患的数据为 N*5000 的数组（N 代表 ictal 和 interictal 的采样数据总条数）。同时，使用 label 数组区分发作期间或者间歇期间的数据，其中发作期间用 1 表示，间歇期间用 0 表示。

2.时域频域转换对比分析

在脑电波数据的分析中，时域分析法主要是通过对脑电波信号的时域波形分析，提取周期和节律等波形特征，作为检测癫痫病发作期的依据[55]。而脑电波信号的频域分析，主要基于傅里叶变换或功率谱估计等，分析脑电信号的功率谱分布及各频段脑电信号的分量[56]。频域分析脑电波信号不仅更为直观，而且更容易观察信号中的噪声，从而进一步剔除噪声。如图 5-9 所示：

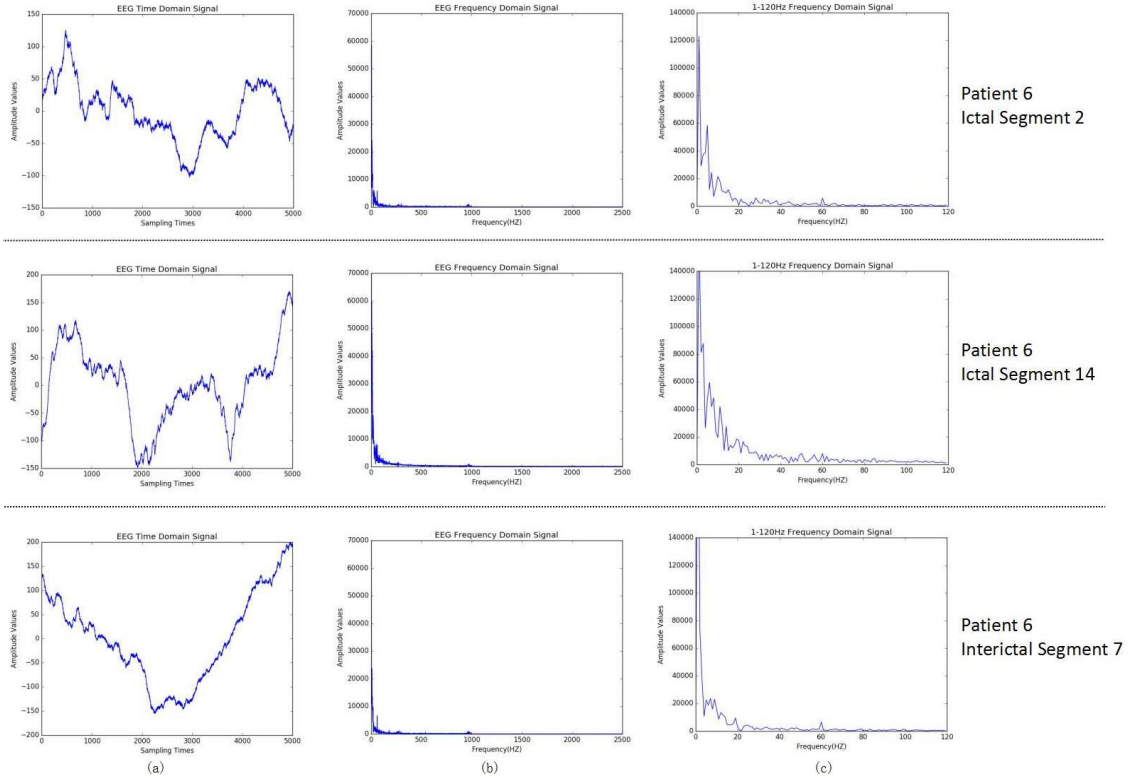


图 5-9 患者 6 的部分时域和频域波形图

图中显示的是癫痫病患者 6 的三个采样数据的波形图，其中(a)列表示时域波形图；(b)列表示的是通过快速傅里叶变换后得到的频域波形图；由于频域波形图列数较高不易观察，因此(c)列显示的是 0-120Hz 的频域波形图。右侧的 ictal 与 interictal 则分别表示的是发作期间和间歇期间的采样脑电波数据。

从图中可以看出，时域波形图较为直观，但是发作期间与间歇期间的脑电波数据并没有大致的区别，特征不明显。而当将时域数据转换成频域之后，可以看出各个频段脑电波的分量，总得看来，随着频率的增加，幅度值大致呈现递减的状态，并且在 150HZ 之后的分量几乎为 0，这很可能是脑电波数据采集过程中因为机械等外界原因产生的不必要噪声，这些噪声信息在时域图形上是无法得到的。(b)列因为频域段跨度较大，数据较集中不便于观察；因此通过(c)列 0~120Hz 的频域段数据可以更为直观的观察出各频段的分量。从(c)列图可以看出，癫痫病患者

发作期间与间歇期间的频域信号有较为明显的差别的。总得来说，发作期间与间歇期间的频域数据相比，各个频段间隔上的抖动变化较频繁，各频段上的分量也更大。

因此，在时域分析和频域分析的对比中可以看到，通过频域分析不仅可以剔除采样数据中存在的噪声数据，还可以更为直观地观察各频段分量的不同。因此采用频域分析法对脑电波进行分析处理更为合适。

3.频域范围选择

从上面的描述也可以得知，在经过傅里叶变换后的频域信号，在 150HZ 之后的频段分量极小，这也就表明在频率为 150HZ 附近及之后的分量极有可能为噪声数据[39]。因此实验中对频段的分量贡献进行叠加处理，并进行绘图展示。如图 5-10 所示：

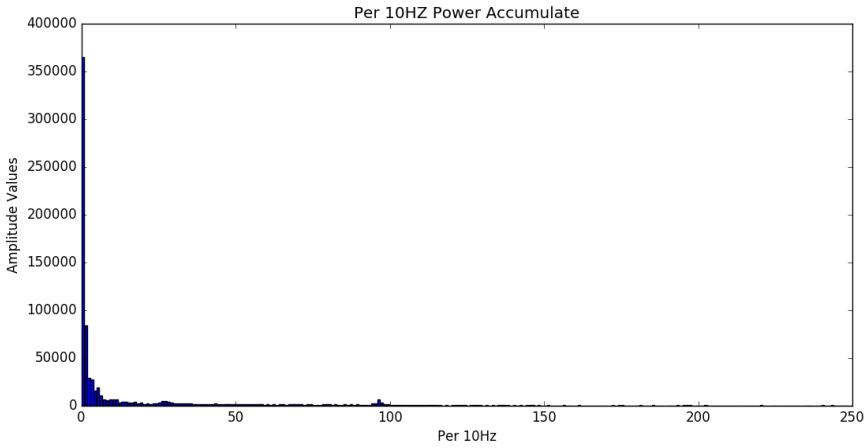


图 5-10 每 10HZ 频域信号叠加统计分析图

图中展示的是频域数据上各频率的分量值，可以看出随着频率的增加，各频率的分量大致呈现下降趋势，也就意味着低频信号对脑电波数据影响较大。

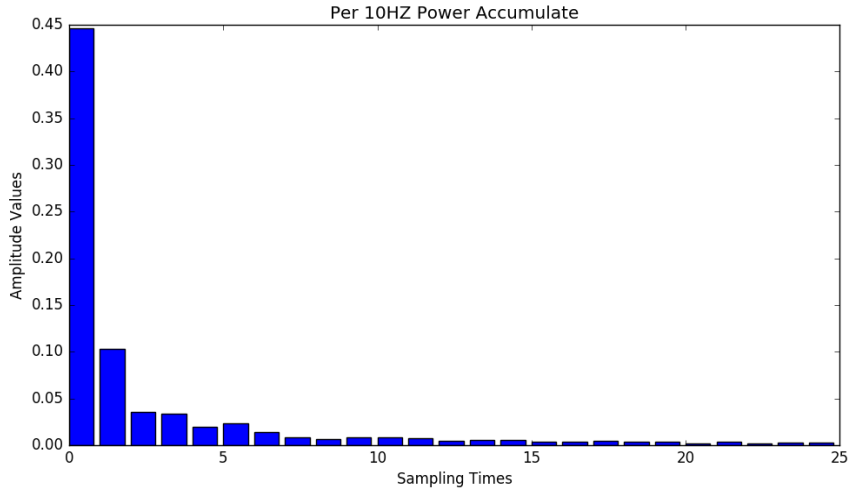


图 5-11 每 10HZ 频域信号叠加统计分析图（1~250HZ）

图 5-11 则展示 0~250HZ 范围内每 10HZ 的分量贡献值，即每 10HZ 为一组，并且计算这一组数据的频率叠加分量总和在整个频域分量总和中的贡献值。图中可以清晰看出，0~10HZ 频率的分量贡献率已经达到了 45%；而随着频率的增加，频域段的贡献值呈现递减趋势，直到 120HZ 趋近于 0。

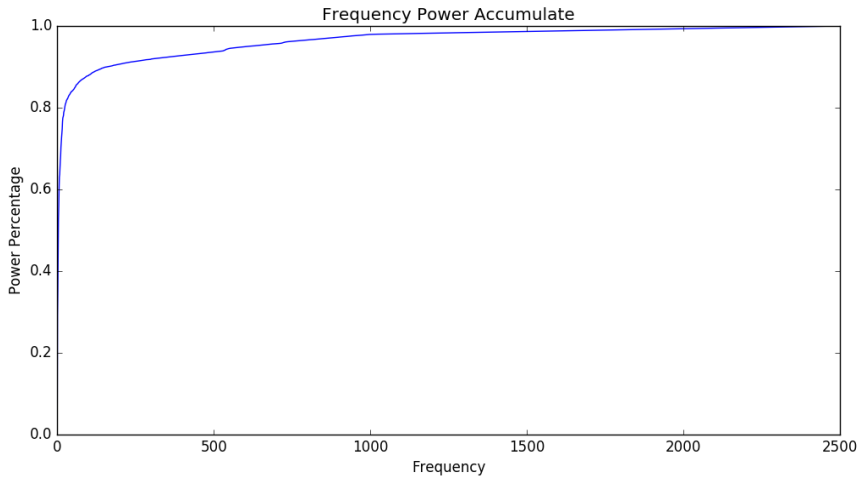


图 5-12 频域能量统计分析图

图 5-12 表明各个频率叠加之后的总体贡献值；例如横坐标 50 则表示：0~50HZ 的频域分量叠加总和对于整个频域分量总和的贡献率。图中清晰的表明：在 150HZ 之前贡献值增长较快，已经接近 90%；而 150HZ 之后，贡献值增长较为平稳。

因此结合以上三图，我们可以确定，0~100HZ 的数据对脑电波数据影响较大，100HZ 之后的频域数据为噪音数据，在之后的分析处理过程中，只需截取 0~100HZ 的频域数据进行数据挖掘即可。

4.对频域数据进行 LLE 降维处理

本实验是采用支持向量机建立分类模型对 EEG 脑电波数据进行分类预测的，因此选择的 0~100HZ 即 100 维频域数据不利于之后预测模型的建立与计算。高维的数据不仅增加预测模型建立的难度，同时也会在计算中消耗大量的时间，从而降低数据挖掘的效率。因此对脑电波频域信号进行降维处理是非常有必要的。在第四章的降维技术研究中表明，使用非线性降维方法中的局部线性降维方法对脑电波频域信号进行降维效果最佳。同时，针对 LLE 算法存在的缺点，第四章第三小节对 LLE 算法进行了相应的改进。本小结的主要内容则为使用改进后的 LLE 算法对脑电波频域信号进行降维处理的实现。

表 5-7 LLE 降维技术的函数示例

Function	Description
KMeansInfo	对整个数据集进行聚类分析
GetInitialNeighbors	获取样本点的初始近邻样本
GetNeighborsMean	获取样本点初始近邻样本的平均距离
GetRealNeighbors	获取样本点的最终近邻样本
RecWeightsMatrix	计算局部重建权值矩阵
GetLowDimens	获取低维空间最佳映射

表 5-7 展示的是改进后的 LLE 降维算法过程涵盖的函数。若设输入数据集为 data, 样本选择近邻个数参考为 neighbors，最终的降维维度为 components, 聚类簇个数为 clusters, 则实现改进的 LLE 降维算法的伪代码如下所示：

```
Algorithm improved_LLE(data,neighbors,components,clusters){
    //使用 KMeansInfo 函数对数据集进行聚类分析
    c = KMeansInfo(data,clusters)
    for(int i=0;i<=data.length;i++)
    {
        //使用 getInitialNeighbor 函数选取样本点的初始近邻样本
        initialNeighbors = getInitialNeighbors(data[i],neighbors);
        //根据初始近邻样本计算平均距离
        mean = getNeighborsMean(initialNeighbors);
        //获取样本点的最终近邻样本
        realNeighbors[i]= getRealNeighbors(initialNeighbors,mean,c[i]);
    }
    //根据 RecWeightsMatrix 函数计算数据集的局部重建权值矩阵
    RWMatrix = RecWeightsMatrix(data,realNeighbors[i]);
    //根据用户输入的 components 以及重建权值矩阵计算最终的低维映射
    result = GetLowDimens(data,RWMatrix,components);
    //返回最终结果 result
    return result;
}
```

首先通过函数 KMeansInfo 对整个脑电波频域数据集进行聚类分析，使得相近或相似的样本处于同一聚类簇中。然后，再通过 GetInitialNeighbors 函数对数据集的每一行即一个样本点选取其初始近邻样本。之后再根据函数 GetNeighborsMean 计算

该样本点的初始近邻样本的平均距离。通过以上的聚类分析和近邻样本的平均距离计算结果，GetRealNeighbors 函数可以得到每个样本点的最终近邻样本。之后再根据传统 LLE 算法的计算思想，首先通过 RecWeightsMatrix 根据每个样本点的最终近邻样本计算数据集的局部重建权值矩阵。最后结合 RecWeightsMatrix 得到低维空间的最佳映射，返回 result 结果即得到数据集的低维表示。

由于输入参数中，样本选择近邻个数参考 neighbors、最终的降维维度为 components、聚类簇个数为 clusters 没有确切的参考输入，因此需要用户通过最终结果仔细调整这一组参数。表 5-8 则展示可在对 Patient5 进行调参时的损失率结果对比：

表 5-8 改进的局部线性嵌入不同参数的性能对比

参数	损失率	AUC 面积
nb=9,cp=7,cl=59	0.163	0.872
nb=10,cp=8,cl=42	0.156	0.859
nb=12,cp=9,cl=30	0.175	0.860
nb=12,cp=6,cl=67	0.154	0.872
nb=14,cp=7,cl=52	0.136	0.906
nb=14,cp=8,cl=47	0.143	0.884
nb=15,cp=7,cl=49	0.161	0.873
nb=15,cp=9,cl=34	0.142	0.897
nb=16,cp=10,cl=22	0.184	0.874
nb=18,cp=7,cl=28	0.153	0.876
nb=18,cp=9,cl=21	0.162	0.864

表中仅为调整参数时的部分结果。第一列代表降维的输入参数，nb 代表的是样本选择的近邻个数, cp 代表的是低维输出的维度，cl 表示数据集进行聚类分析时最终簇的个数。第二列代表数据集低维输出时的损失率。第三列表示通过 SVM 分类后的 AUC 面积。从表中可以看出，输入参数会在一定程度上影响着降维结果以及最后的分类效果；当 nb=14，cp=7，cl=52 时降维后的损失率和分类的 AUC 面积值是最优的，因此选择这组参数进行降维。当然，针对不同的癫痫病患者的脑电波数据也会有不同的参数输入，由于数据较多这里便不进行一一展示。

5.4 结果展示与分析

通过以上的系统分析与设计实现，本系统主要实现了“人口死亡”数据集处理和“癫痫病脑电波”数据集处理。

（1）“人口死亡”数据集预处理

在“人口死亡”数据集预处理中依次对数据集进行了数据集成、异常数据处理、数据的特征选择、数据变换以及最后的缺失值填补。在填补缺失值这一环节使用改进 SMOTE 算法合成新的少数类样本，从而达到改善非平衡数据集的目的；之后再通过随机森林建立分类模型对数据集中“死亡方式”这一属性的缺失值进行填补，缺失值填补的性能分析如图 5-13 所示：

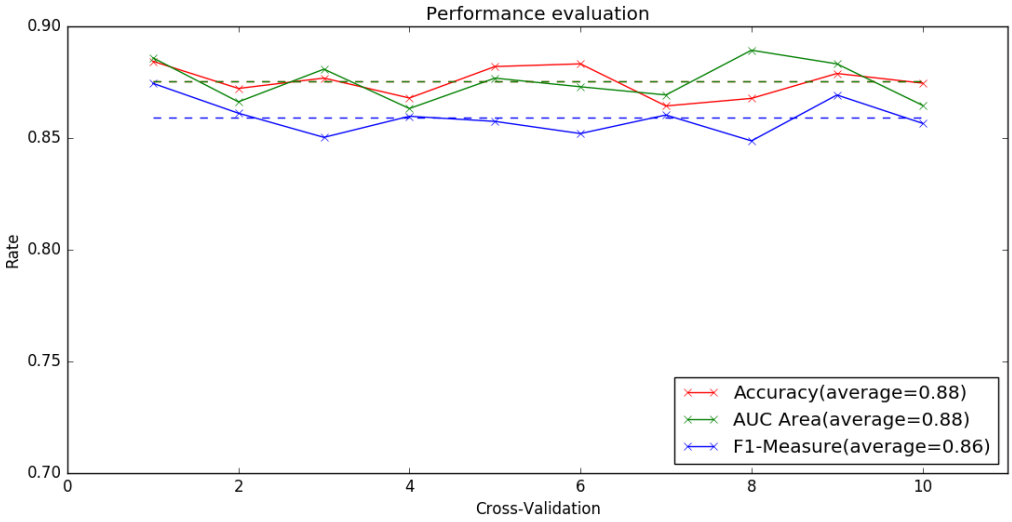


图 5-13 cv=10 时的分类结果性能对比图

从图中可以看出，分类精度和 AUC 面积的均值可以达到 0.88，F1-Measure 的均值为 0.86。从这三项性能评估指标可知，缺失值填补分类模型建立较为成功，能够较准确地对“死亡方式”缺失值进行填补。

（2）“癫痫病脑电波”数据集预处理

而在对“癫痫病脑电波“数据集进行预处理以及最后的 SVM 分类处理后，最后的分类预测结果如图 5-14 所示：

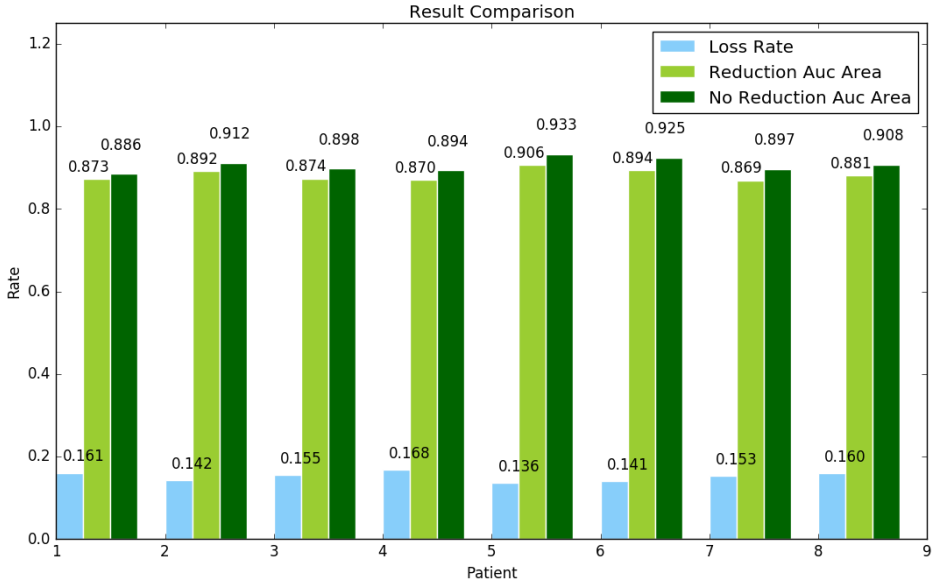


图 5-14 10 名患者的处理结果性能对比图

从图中可以看到，在使用改进的 LLE 降维之后，8 名癫痫病患者的降维损失率都控制在 0.13~0.17 之间；而通过降维之后再通过 SVM 建立预测模型后的重要性能指标 AUC 面积值在 0.86~0.91 之间，比不使用降维技术的预测模型的 AUC 面积值平均减少 2-3 个百分比，差距在可接受范围之内。而使用 LLE 降维后与未降维时的处理分析消耗时间对比如图 5-15 所示：

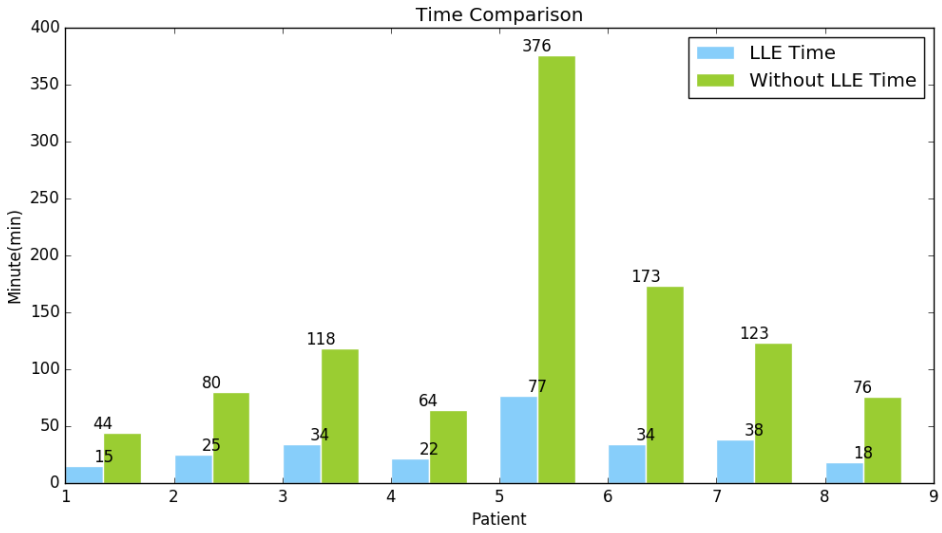


图 5-15 频域信号降维与不降维的处理时间对比图

图中可以看到，在使用 LLE 降维处理脑电波数据之后，分析处理的时间消耗大大地减少了，几乎只有未降维处理时的五分之一。这也就表明降维方法不仅可以保障挖掘质量，还能有效地提高数据挖掘的效率。

5.5 本章小结

本章介绍了人口死亡数据集和癫痫病脑电波数据集预处理阶段的需求以及总体框架设计，并对架构的核心部分进行了详细阐述和结果演示。通过结果表明，两个预处理模块的架构设计正确并可以达到良好的预处理效果，在提高数据挖掘效率及挖掘质量方面有明显的效果。

第六章 全文总结与展望

6.1 全文总结

医疗领域是一个面向全民提供服务的重要领域，与民生息息相关。而随着信息技术的发展，医疗数据正在以惊人的速度增长；如何从医疗健康数据中挖掘出有价值的信息成为当今的一个探讨热点。而在数据挖掘中，由于医疗数据所具有的独有特点，对医疗数据进行数据预处理分析是必要的。

本文研究数据挖掘中数据预处理的常用关键技术，并将大数据预处理技术应用于实际的数据集预处理分析中。针对健康数据集的特点，分别使用了不同的预处理方法，并在某些方法技术上提出了具有创新意义的改进方法。具体的研究贡献如下：

（1）“人口死亡”数据集缺失值填补方法的改进。在此数据集中，通过各个现代填补缺失值方法的对比分析进而选择随机森林算法。而面对该数据集独有的非平衡性问题时，通过过采样技术 SMOTE 算法进行非平衡性改善；同时针对 SMOTE 算法存在的问题，提出一种基于重心的 SMOTE 改进方法。通过实验对比分析，改进的 SMOTE 算法可以更好的解决边缘数据和噪声数据在合成新样本时产生的偏差；在改善数据集平衡性时效果更佳。

（2）“癫痫病脑电波”数据集频域信号降维方法的改进。在对脑电波信号进行预处理时，采用局部线性嵌入算法对频域数据进行有效降维提高数据挖掘的效率；同时针对该算法存在的缺陷提出一种基于 K-Means 和均值的邻域点选择方法，通过类簇和均值的限制，样本点可以自适应的寻找合适邻域点。通过实验对比分析表明，改进后的局部线性嵌入算法有效地提高了脑电波频域信号的降维效果。

（3）预处理分析技术的具体应用的实现。对“人口死亡”数据集和“癫痫病脑电波”数据集进行了预处理分析实现；将预处理相关技术应用于具体实践中，为下一步的数据挖掘工作提供有效的高质量数据集。同时，改进的 SMOTE 算法和局部线性嵌入降维算法不受具体的数据集限制，可推广至其他应用中。

总之，全文对两个具体数据集进行了预处理分析与实现，并针对相关问题和缺陷提出创新改进，有效提高了数据集挖掘时的准确性和效率。

6.2 未来展望

虽然国家在大力推行国家医疗信息化的建设，但由于医疗领域与信息技术领

域的巨大差异性使得进程发展较为缓慢。由于数据预处理分析的相关技术、理论非常多；而数据挖掘在医疗信息领域的应用实例较少，因此造成一些内容的研究分析还不够全面和深入，因此在接下来的研究工作中，还需要着重注意以下几个方面：

（1）在将数据挖掘技术应用于医疗数据时，应该多了解医疗数据的独有特点。针对不同数据的诊断特点制定独有的数据预处理及数据挖掘方案，从而有效提高数据挖掘质量和效率。

（2）通过使用改进的 SMOTE 算法可以合成更好的新样本，从而达到改善非平衡数据集的目的。由于改进后的 SMOTE 算法每次只合成 2 个样本点，因此时间消耗较大；同时，由边缘性样本和噪声样本合成的新样本偏离样本中心的情况并未能完全消除。因此，需进一步研究改善非平衡数据集的可行性方法。

（3）改进后的局部线性嵌入算法依然存在无法确定最优参数的问题。实际应用中，使用局部线性嵌入算法时的参数输入只能通过实验的结果输出不断调整；因此实验的结果不一定是全局最优，只能保证局部最优。

（4）本文只着重研究了随机森林相关算法和局部线性嵌入降维算法，而数据预处理领域的方法非常多，因此接下来需要研究更多其他的方法并针对其特点进行相关改进，将其真正应用到医疗领域中。

致　谢

花样年华之时与电子科技大学相遇，接下来的时光里，一直美好。回首三年的求学历程，对于那些不离不弃陪伴我的人，我心中充满感激。感谢电子科技大学赠予我的一切：酸甜苦辣的经历到最后都会沉淀为一杯人生甘醇的美酒。

首先，要感谢我的研究生导师卢光辉副教授。在充实的三年研究生期间，卢老师做学问时执着追求的态度和为人处事的平和大度都是我今后工作、生活学习的榜样。在学习和生活中，卢老师多次给予我建设性的意见，使我如沐春风，豁然开朗。在此谨向卢老师表示我最诚挚的敬意和感谢！

其次，我要感谢闫朝利、余宽、刘璐、陈琪、杨中元同学，以及同门们。在学习上，大家虽是竞争对手但更是朋友，期间大家共同交流、学习进步。生活中，是你们在我失落时给我鼓励，在我无助时给我帮助，让我在荆棘路上不断前行。一路风雨，一路坎坷，感谢大家一路的陪伴，让我的研究生生活充满了乐趣与温暖。愿岁月不老，友谊长存。

最后感谢我最亲爱的父母和兄弟姐妹们。感谢父母的养育和家人不求回报的无限支持。是你们给了我一个温暖和谐的家，给我铸造了一个坚实而温馨的避风港。在未来的日子里，我会更加努力，不辜负你们对我的殷殷期望！

祝愿我生活三年的母校蒸蒸日上，欣欣向荣。

参考文献

[1] 石思优. 基于主题模型的医疗数据挖掘研究[D]. 广东技术师范学院,2015

[2] 马灿. 国内外医疗大数据资源共享比较研究[J]. 情报资料工作,2016,(03):63-67

[3] 曹德贤. 寄语“十二五”医疗卫生信息化[J]. 中国数字医学,2011,(01):9

[4] Han J, Pei J, Kamber M. Data mining: concepts and techniques[M]. Elsevier, 2011

[5] 李惠先,封二英. 大数据时代医学研究面临的机遇与挑战[J]. 计算机光盘软件与应用,2014,(23):138-139

[6] Kop R, Hoogendoorn M, ten Teije A, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records[J]. Computers in biology and medicine, 2016, 76: 30-38

[7] Marques F J, Moutinho A, Vieira S M, et al. Preprocessing of Clinical Databases to improve classification accuracy of patient diagnosis[J]. IFAC Proceedings Volumes, 2011, 44(1): 14121-14126

[8] 林予松,王培培,刘炜,李润知,王宗敏. 医疗体检数据预处理方法研究[J]. 计算机应用研究,2017,(04):1-5

[9] Hoshyar A N, Al-Jumaily A, Hoshyar A N. The Beneficial Techniques in Preprocessing Step of Skin Cancer Detection system Comparing[J]. Procedia Computer Science, 2014, 42: 25-31

[10] 白凤伟.数据预处理系统的几个关键技术研究与实现[D]. 北京交通大学,2012

[11] Croy Calvin D,Novins Douglas K.Imputing Missing Data[J].Journal of the American Academy of Child &Adolescent Psychiatry.2004;43(4):380-381

[12] Clarke B, Fokoue E, Zhang H H. Principles and theory for data mining and machine learning[M]. Springer Science & Business Media, 2009

[13] Fodor I K. A survey of dimension reduction techniques[J]. 2002

[14] GILLES J. Empirical Wavelet Transform[J]. IEEE Transactions on Signal Processing, 2013, 61(16): 3999-4010

[15] B.Principal component analysis in linear systems: Controllability, observability, and model reduction. Automatic Control on IEEE Transactions, 1981, 26(1):17-32

[16] Baldacci L., M. Golfarelli, A. Lumini, et al. Clustering techniques for protein surfaces. Pattern Recognition, 2006, 39(12): 2370-2382

[17] Gupta A, Mehrotra K, Mohan C. A clustering based discretization for supervised leaming[J]. Statisticsand Probability Letters, 2009, 80(9-10):816-824

[18] Kerber R.Chimerge: discretization of numeric attributes[C]. Proceedings of Ninth National Conference on Artificial Intelligence, San Jose, California, AAAI Press. 1992; 123-128

[19] 刁树民,于忠清.概念分层在人口普查数据中的应用[J]. 现代电子技术,2006,(20):47-49

[20] 李圣瑜.调查数据缺失值的多重插补研究[D]. 河北经贸大学,2015

[21] Cover T M. Rates of convergence for nearest neighbor procedures[C]. Proceedings of the Hawaii International Conference on Systems Sciences. 1968: 413-415

[22] Kim C, Hwang K B. Naive Bayes classifier learning with feature selection for spam detection in social bookmarking[J]. ECML PKDD Discovery Challenge 2008, 2008: 32

[23] Apte C, Damerau F, Weiss S. Text mining with decision rules and decision trees[M]. IBM Thomas J. Watson Research Division, 1998

[24] T.Joachims. Text categorization with support vector machines:Learning With Many Relevant Features.In Proceedings of 10th European Conference on Machine Learning,1998:137-142

[25] C.Hsu, C.Lin. A comparison on methods for multi-class support vector machines, IEEE Transactions on Neural Networks.2002,13:415-425

[26] E.Wiener. A neural network approach to topic spotting. In Proceeding of the 4th Annual Symposium on Document Analysis and Information Retrieval(SDAIR95),Las Vegas,NV,1995

[27] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C]. Icml. 1996, 96: 148-156

[28] Anzai Y. Pattern Recognition & Machine Learning[M]. Elsevier, 2012

[29] Breiman L. Random forests [J]. Machine learning, 2001, 45(1): 5-32

[30] 贺捷. 随机森林在文本分类中的应用[D]. 华南理工大学,2015

[31] Ho T K. The random subspace method for constructing decision forests[J]. IEEE transactions on pattern analysis and machine intelligence, 1998, 20(8): 832-844

[32] Weiss G M. Mining with rarity: a unifying framework[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 7-19

[33] Schapire R E. Pattern Languages are not Learnable[C]. COLT. 1990: 122-129

[34] Schapire R E. The strength of weak learnability[J]. Machine learning, 1990, 5(2): 197-227

[35] Bache K, Lichman M. UCI machine learning repository[J]. School of Information and Computer Science, University of California, Irvine, CA

[36] Maratea A, Petrosino A, Manzo M. Adjusted F-measure and Kernel Scaling for Imbalanced Data Learning[J]. Information Sciences, 2013

[37] Tay F E H, Shen L. A modified chi2 algorithm for discretization[J]. IEEE Transactions on knowledge and data engineering, 2002, 14(3): 666-670

[38] Wang J. Geometric structure of high-dimensional data and dimensionality reduction[M]. Springer Berlin Heidelberg, 2011

[39] Yin Z, Zhang J. Identification of temporal variations in mental workload using locally-linear-embedding-based EEG feature reduction and support-vector-machine-based clustering and classification techniques[J]. Computer methods and programs in biomedicine, 2014, 115(3): 119-134

[40] He X, Cai D, Han J. Learning a maximum margin subspace for image retrieval[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(2): 189-201

[41] Pearson K. LIII. On lines and planes of closest fit to systems of points in space[J]. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901, 2(11): 559-572

[42] Haeb-Umbach R,Ney H.Linear discriminant analysis for improved large vocabulary continuous speech recognition.Acoustics,Speech,and Signal Processing.International Conference on IEEE,1992,1:13-16

[43] Kruskal J B. Nonmetric multidimensional scaling: a numerical method[J]. Psychometrika, 1964, 29(2): 115-129

[44] 王靖. 流形学习的理论与方法研究[D]. 浙江大学,2006

[45] 龚铁梁. 数据降维算法研究及其应用[D]. 湖北大学,2012

[46] Tenenbaum J B. The Isomap Algorithm and Topological Stability[J]. Science, 2002, 295(5552): 7a

[47] Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. science, 2000, 290(5500): 2319-2323

[48] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural computation, 2003, 15(6): 1373-1396

[49] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326

[50] 王勇. 局部线性嵌入算法的稳健性改进及有监督扩展[D]. 国防科学技术大学,2007

[51] Saul L K, Roweis S T. Think globally, fit locally: unsupervised learning of low dimensional manifolds[J]. Journal of Machine Learning Research, 2003, 4(Jun): 119-155

[52] De Ridder D, Duin R P W. Locally linear embedding for classification[J]. Pattern Recognition Group, Dept. of Imaging Science & Technology, Delft University of Technology, Delft, The Netherlands, Tech. Rep. PH-2002-01, 2002: 1-12

[53] Kouropteva O, Okun O, Pietikäinen M. Selection of the optimal parameter value for the locally linear embedding algorithm[J]. FSKD, 2002, 2

[54] Hong Y, Kwong S. Learning assignment order of instances for the constrained K-means clustering algorithm[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, 39(2): 568-574

[55] 袁琦. 癫痫脑电的分类识别及自动检测方法研究[D]. 山东大学,2014

[56] Lee G, Kwon M, Kavuri S, et al. Action-perception cycle learning for incremental emotion recognition in a movie clip using 3D fuzzy GIST based on visual and EEG signals[J]. Integrated Computer-Aided Engineering, 2014, 21(3): 295-310

攻读硕士学位期间取得的成果