# 本科毕业论文（设计）

# 基于自编码器聚类算法的心脏功能评级算法研究

# RESEARCH ON HEART FUNCTION RATING BASED ON AUTOENCODER CLUSTERING ALGORITHM

石博文

哈尔滨工业大学

2024 年 5 月

密级：公开

# 本科毕业论文（设计）

# 基于自编码器聚类算法的心脏功能评级算法研究


本　　科　　生：石博文

学　　　　号：120L020629

指　导　教　师：袁永峰教授

专　　　　业：计算机科学与技术

学　　　　院：计算学部


答　辩　日　期：2024 年 5 月 26 日


学　　　　校：哈尔滨工业大学

# 摘　要

科学研究表明实验室测量指标、肝脏功能和常规超声指标与心功能评级有着十分密切的联系，但具体的影响程度以及关系尚不明确。本课题基于哈尔滨医科大学提供的脱敏后患者相关体检数据，使用了拓扑聚类算法和自编码器深度聚类算法对患者样本进行聚类，并从聚类结果的 K - S 检验和聚类结果的医学合理性两个角度，对比聚类结果的优劣并确定效果最好的聚类算法。本文工作首先通过随机森林算法，以及数据滤波操作对实验数据集进行了数据维度删除以及缺失数据填补；随后实现了拓扑聚类算法，构造了两种数据升维算法，三种不同网络结构的自编码器，并且探究了聚类中心初始化个数对聚类结果的影响。最终经过实验结果对比和理论分析，确定了采用线性插值算法进行数据升维，初始化聚类中心个数为 3，自编码器网络结构为 Conv2DTranspose 的卷积神经网络能够得到最好的聚类结果。

基于该自编码器模型，实验数据被分为了三个等级。经过 K - S 检验，这三类数据两两之间的强有效区分（即检验 p - value 值小于 0.01）维度数量分别为 9、4、6 维，且有效区分数量均在十个以上。其中维度 INR 凝血酶原时间和肝脏联合弹性指标 LFI（Med）在三个类别间的双样本 K - S 检验中都属于有效区分维度，并且通过对医学合理性的分析发现三个聚类类别的患者样本心功能存在差异，表明该模型能够合理有效地通过患者的体检数据对其心功能进行评级。

**关键词**:心功能评级；自编码器算法；深度聚类算法；卷积神经网络算法

# Abstract

Scientific research shows that laboratory measurement indicators, liver function and conventional ultrasound indicators are very closely related to cardiac function ratings, but the specific impact and relationship are still unclear. This topic is based on the patient-related physical examination data provided by Harbin Medical University. It uses topological clustering algorithm and autoencoder deep clustering algorithm to cluster patient samples, and performs the K-S test on the clustering results and the medical evaluation of the clustering results. From the two perspectives of rationality, compare the advantages and disadvantages of clustering results and determine the best clustering algorithm. This work first used the random forest algorithm and data filtering operations to delete data dimensions and fill in missing data on the experimental data set; then implemented a topological clustering algorithm, constructed two data dimensionality increasing algorithms, and constructed three different network structures of autoencoders, and explored the impact of the initial number of cluster centers on the clustering results. Finally, through comparison of experimental results and theoretical analysis, it was determined that the convolutional neural network using linear interpolation algorithm to increase the data dimension, initializing the number of cluster centers to 3, and the autoencoder network structure Conv2DTranspose can obtain the best clustering results.

Based on this autoencoder model, the experimental data is divided into three levels. After K-S test, the number of strong and effective distinctions between these three types of data (that is, the test p-value value is less than 0.01) is 9, 4, and 6 dimensions respectively, and the number of effective distinctions is more than ten. Among them, the dimensions INR and LFI (Med) are effective distinguishing dimensions in the two-sample K-S test between the three categories, and through the analysis of medical rationality, it was found that the three clustering categories There are differences in cardiac function among patient samples, indicating that the model can reasonably and effectively rate patients' cardiac function through their physical examination data.

**Keywords:** cardiac function rating， autoencoder algorithm， deep clustering algorithm，convolutional neural network

# 目　录

# 第 1 章　绪　论

## 1.1　课题来源

本课题来源于国家自然科学基金项目“动态环境下狭窄腔道手术机器人感知与控制研究”。科学研究表明肝脏功能在一定程度上影响着心脏功能[1]，常规超声指标与心功能评级有着十分密切的联系[2]。除此之外，血常规生化指标以及腹部栓塞等对心脏功能也有着潜在的影响。那么在这诸多影响因素的作用下，是否可以通过聚类学习方法，根据心、肝、血液等的生化指标对病人进行心功能评级是当前临床医学研究的一项重点问题。本课题的中心思想在于，基于肝脏、心脏超声指标，以及其他包括但不限于直接胆红素、凝血酶原活动度等生物化学指标，构建基于深度学习方法的聚类分析模型，挖掘讨论引起心脏功能变化的诸多指标，进行对比学习得出心功能评级，衡量心脏的状况，并且根据聚类结果优化聚类网络结构。

## 1.2　研究目的以及意义

本课题旨在：基于肝脏、心脏超声指标，以及其他包括但不限于直接胆红素、凝血酶原活动度等生理指标，构建一个以深度学习为方法的，使用不同网络的聚类分析模型，探寻让心脏功能或变好或变坏的相关指标，进行对比学习得出心功能评级，通过病人的体检数据衡量病人心脏的状况，并且根据评级高低给出相关的医学和生活建议。

课题研究的意义如下：

1） 综合评估各项指标对心脏功能进行分级，探讨不同指标对心脏功能是否有影响以及其具体作用程度如何，可辅助判断患者心脏问题的根源及改善方向，为医生的治疗方案提供建议，对心脏疾病的防治具有重要价值；

2） 利用深度学习以及神经网络相关方法探索心脏功能评分的有关研究任务较少，本课题的研究有一定创新性意义，可以为未来的研究提供参考；

3） 对于未来维度更多的数据，可以采用与本课题类似的方法进行研究；或者可能有类似的数据可以在现有研究的基础上进行推广，提供一种可能的解决方案，即利用已知的评级数据提供另一组相关的数据评级，通过扩展数据集的方式对网络进行新一轮的训练，以达到课题研究结果复用的效果和目标。

## 1.3 国内外在该方向上研究现状以及分析

心脏与肝功能的密切关系已被众多医学研究证实[3]，中医、西医等诸多观点也充分展示了心肝相生相克的密不可分的关系[4]。血液中某些蛋白质和酶的含量也有影响心脏功能的可能性，如谷氨酰转肽酶，可加速心脏功能的恢复[5]，但目前的研究只说明了这些指标的变化会对心脏功能产生影响，但是没有方法论证这些指标是否都会对心脏功能产生影响，并且也不清楚其作用的程度。因此，有必要筛查各因素具体的影响程度。

近年来，将机器学习方法用于疾病研究的研究也很多。参考文献[6]总结了近年来各种机器学习方法用于疾病预测的研究；Jeffrey S Bennett[7]利用 K 均值聚类方法将层粘连蛋白数据样本聚类为 3 个簇，并使用卡方检验进行显著性统计分析；Jabir Al Nahian 等[8]分别利用随机森林、SVM、逻辑回归和 KNN 方法对人类常见疾病进行预测，在其数据集上平均准确率为 76.45%。机器学习在疾病分类方面也取得了很大的成就：Ahmed Zriqat 等[9]利用朴素贝叶斯分类器和支持向量机（SVM）模型，利用数据挖掘分类方法根据患者的症状判断心脏病的类型，在交叉验证数据集上，两种方法的准确率分别达到 78.88%和 76.57%；Fajr Ibrahem Alarsan 和 Fajr Ibrahem Alarsan [10]使用决策树和梯度增强树算法分析心跳特征并对心脏病进行分类。当使用随机森林时，最佳分类准确率达到 96.31%，可以划分出四种不同的心跳类型。

近年来，利用机器学习方法对血液指标进行分析的研究呈现出日益增长的趋势。例如，采用软边缘支持向量机（SVM）模型[11]，以及对血液生化成分指进行筛选，已经在识别影响非酒精性脂肪性肝病的因素方面取得了一些重要成果。这些研究发现了一些显著的影响因素，对于本课题的研究具有重要的借鉴意义。

然而，当前尚未见到利用深度学习聚类分析方法对心脏、肝脏、血液生化成分等综合因素进行心脏功能评分的研究。因此，开展这一课题具有重要的开拓意义。通过综合考虑多种因素，包括心脏功能、肝脏健康以及血液生化成分等，利用深度学习聚类分析方法进行心脏功能评分，可以为心血管疾病的预防、诊断和治疗提供更全面的信息和更准确的评估。这不仅有助于优化医疗资源的分配，还能够提高患者的治疗效果和生活质量。因此，开展该课题具有重要的理论和实践意义，有望为医学领域的进步和人类健康的促进做出贡献。

## 1.4 课题研究流程以及论文结构

本课题使用深度学习聚类分析的方法进行心功能评级，主要研究的内容如下：

1） 基于数据给出的心脏、肝脏、以及血液生化成分指标，初步筛选出会影响心脏功能以及影响心功能评级的指标；

2） 构建多种聚类分析模型，包括拓扑聚类算法[12]和自编码器深度聚类算法[13]，使用不同的网络结构进行深度学习算法的训练，得到多种不同的聚类结果，计算和心脏功能相关指标的标准差，用于后续对结果的评价和分析；

3） 根据聚类分析的实验结果以及给出的每个数据的评级标签，对未分级样本的心功能进行分类评级；

4） 对于实验结果，结合医学知识以及聚类结果相关指标，如 KS 散度，对聚类结果的合理性和有效性进行评价，并比较不同聚类模型之间的优劣。

根据研究流程，论文各章内容简要概括如下：

第一章为绪论部分。主要介绍了课题来源，课题研究的目的和意义，国内外在该方向上的研究现状以及分析，以及简要阐述课题的研究流程和论文的行文结构。

第二章为介绍数据预处理过程。通过对哈尔滨医科大学提供的脱敏后数据进行观察、计算和分析，使用滤波函数方法，初步筛选出一些对聚类算法影响显著和影响低下的数据维度，进行数据维度的删除；同时计算数据的缺失率，针对包含缺失数据的维度采取相应的算法进行数据填补。

第三章为拓扑数据分析聚类方法的实现。通过拓扑数据分析聚类和 Mapper 算法[14]，通过高维开方体对数据进行覆盖，得到患者心脏功能状况的拓扑图，其中情况相似的患者在图中通过边产生链接，尤为相似的患者聚集成一个拓扑图中的点，通过拓扑图中点与点之间的联通关系能够将患者样本聚集为不同的类别并打标签。

第四章为自编码器深度聚类方法的实现。通过构建一个自编码器，以输入数据为目标，使经过神经网络的输出数据尽可能地和输入数据相似，并以此为目标调整网络层中的参数，在达到一定低程度的损失之后，提取隐藏层的信息作为聚类层的数据数据进行聚类层的训练；随后构建聚类层，将输入特征转化为聚类标签概率，概率计算由 t-分布给出，度量了样本点和样本中心的相似性，通过不断的迭代训练，使预测分布不断迫近构造目标分布，使

得两个分布的相似性达到一定临界值，停止训练，最终得到患者样本的聚类标签。

第五章为聚类方法的结果对比。通过对比拓扑分析技术和自编码器深度聚类技术，以及自编码器深度聚类在使用不同网络结构以及不同数据升维方法后得到的不同聚类结果，分析产生结果的原因以及选出最优的聚类方法。

# 第 2 章　心肝相关体检数据预处理

## 2.1　初步观察数据

本课题使用哈尔滨医科大学提供的患者体检数据，数据中除去了患者的详细身份信息，仅保留了患者的体检相关指标进行脱敏，不会泄露患者隐私，在本文接下来提到的数据均为此脱敏后数据。数据维度为 208×106：其中包含了病人的体貌特征例如性别身高体重，血液生化指标例如谷丙转氨酶等等。经过观察，可以发现数据中含有一些数据特征相关的维度，例如 VFM-能量损耗中位数、平均值等，高铝祷告相关性，需要使用滤波函数进行数据删除操作。

除此之外，数据中还有部分病人资料不完整，存在数据缺失的问题，因此考虑使用随机森林的方法对数据进行数据填补。

## 2.2　数据填补

考虑到数据中有些维度信息为文字信息，例如 Child-Pugh 分级维度的信息为 ABC 评级，考虑到本课题的模型仅使用数值型数据，因此在数据中删除了 Child-Pugh 维度。随后计算各个维度数据的缺失率，情况如下表所示：

表 2-1 含有缺失数据维度

| 数据维度名称 | 缺失率/% |
|---|---|
| PT 凝血酶原时间 | 5.29 |
| PTA/PT%凝血酶原活动度 | 6.25 |
| INR 凝血酶原时间 | 5.29 |
| 肝脏联合弹性指标（含 E、LFI 等） | 6.25 |
| VFM 能量损耗 | 7.21 |
| TP 总蛋白 | 2.40 |
| 直接胆红素 | 2.40 |
| MELD 分级 | 4.80 |
| 中性粒细胞 | 4.32 |
| SCr 血肌酐 | 3.36 |
| 肝硬化程度 | 5.76 |
| 总缺失率 | 20.19 |

经过观察，发现 VFM-能量损耗相关指标中部分维度缺失率相同，为某些样本个体缺少相关整体体检指标，因此在表格中成整体显示，其中部分数据缺失率超过 5%，缺失率较高，接下来使用随机森林[15]方法对数据进行填补。

随机森林的算法核心思想是通过将含有缺失值的数据维度序列分为两部分，缺失部分作为预测部分，完整部分作为训练部分，通过利用其它数据完整的样本信息和对应标签，通过回归训练，给出缺失部分应该填补的标签值。算法完整流程伪代码如下所示：

**算法 1** 随机森林算法填补缺失值

**输入：**含有缺失值的数据矩阵 Matrix

**输出：**完整的数据矩阵 $Matrix^{fill}$

**1：**提取含有缺失值的维度列表 Missing

**2：For** Feature **in** Missing **do:**

**3：** 提取 Feature 中完整数据为 X_train，标签为 Y_train

**4：** 提取 Feature 中缺失数据为 X_test，标签为 Y_test

**5：** 实例化随机森林回归器 RandomForestRegressor

**6：** 进行回归训练

**7：** 使用训练后的回归器预测缺失值并填补进 $Matrix^{fill}$

**8：**检查是否由数据缺失

**9：return** $Matrix^{fill}$

在应用了上述算法之后，最终得到了一份维度为 208×105 的不含缺失数据的数值化样本。

## 2.3 数据删除

数据删除这一小节的任务是通过使用低方差滤波和高相关滤波两种滤波方法，删除一些相关性较强以及变化幅度极小的数据维度，认为这些维度对聚类结果的贡献微乎其微，为了提高训练的速度，予以删除。

在进行相关滤波操作之前，首先对数据进行 min-max 归一化处理，这样操作的目的是将数据值映射到 0 到 1 的闭区间中，公式如下，（其中 data 表示上述填补后的数据，i 表示某一样本行）：

$$data[i]=\frac{data[i]-\min(data)}{\max(data)-\min(data)} \tag{2-1}$$

随后计算归一化后的数据的方差，在此认为数据方差低于 0.01 的列维度变量变

化幅度不大，对聚类模型的构建贡献较少，因此舍弃这些维度。

随后进行高相关滤波的操作，同样使用经过归一化后的数据，计算两两维度之间的 Pearson 相关系数，公式如下（其中 X，Y 表示两个维度）：

$$\rho_{X,Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \tag{2-2}$$

通过可视化函数库显示各个维度之间的数据相关性，如下图所示：

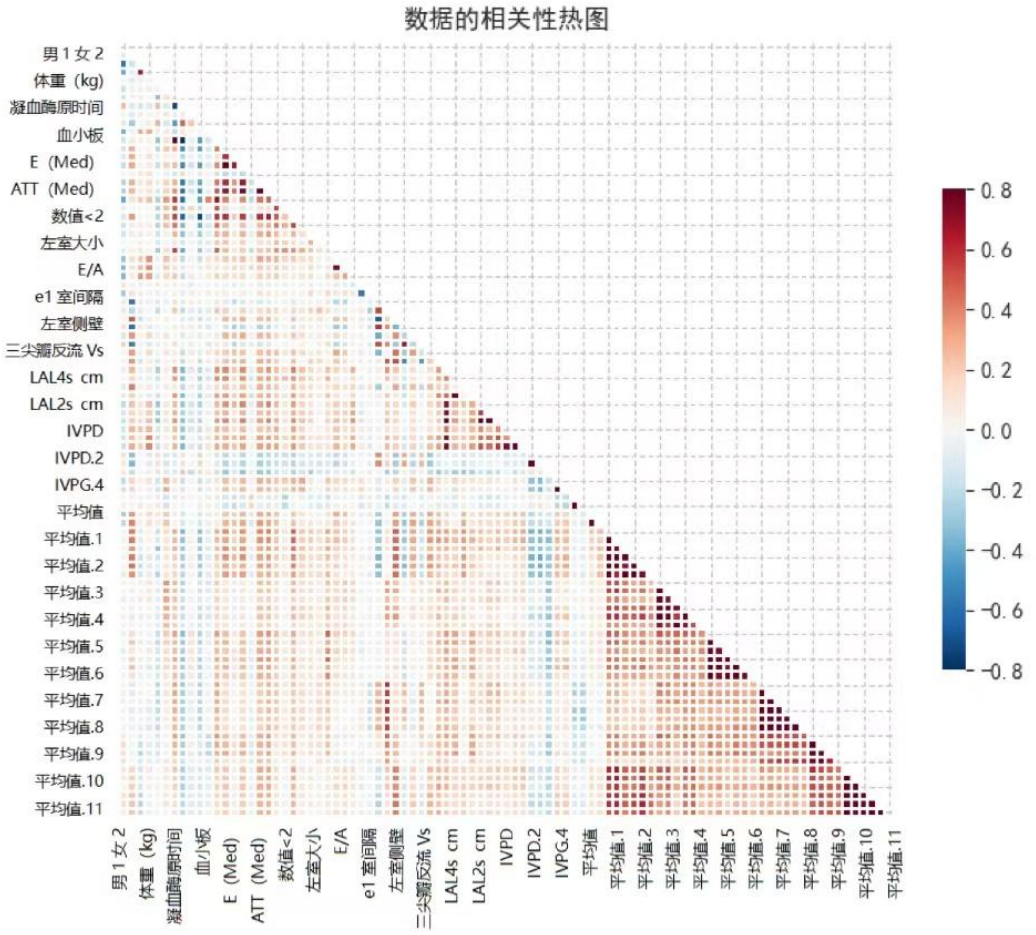


图 2-1 维度相关性指标

在此认为两个为度之间的 Pearson 系数大于 0.85，两个维度就具有较高的相关性。但是在审查这些相关性较高的维度的时候，发现有一些数据维度来自于同一体检检测指标的不同数据特征，呈现两两相似的情形，例如 VFM-能量损耗中等容收缩期 AVE 重点考虑体表面积标化的平均值、中位数和最大值。为此，为了保证不丢失数据信息，将这些三个两两相似，但是统一于同一个指标大类名称的维度，删除其中的两个，保留其中的一个，作为初步筛选的数据维

度保留在实验数据中。最终得到的删除了的维度信息如下表所示：

表 2-2 舍弃的维度名称

| 过滤的维度名称 | | | | |
|---|---|---|---|---|
| DBIL 直接胆红素 | 实验室 APRI | 正常 0 肝病 1 | 中位数 | 平均值.6 |
| ALP 碱性磷酸酶 | IVPG.1 | 左房容积指数 | 最大值 | 平均值.6 |
| ALT 谷丙转氨酶 | IVPD.3 | LAL4s/cm | 平均值.1 | 中位数.6 |
| ALB 白蛋白 | IVPG.3 | LAS4s/cm | 平均值.2 | 最大值.7 |
| TBIL 总胆红素 | 中位数.8 | IVPD | 中位数.2 | 平均值.8 |
| AST 谷草转氨酶 | 最大值.8 | IVPD.2 | 平均值.4 | 平均值.9 |
| Aindex(Med) | PT 凝血酶原时间 | IVPD.4 | 中位数.4 | 平均值.10 |
| Child-Pugh 分级 | PTA/PT% | IVPD.5 | 最大值.4 | 平均值.10 |
| | 凝血酶原活动度 | | | |
| 实验室 Fib-4 | E(Med) | 平均值 | 平均值.5 | 中位数.11 |

## 2.4 本章小结

经过本章的工作，对哈尔滨医科大学提供的，脱敏后的患者心功能相关体检数据进行了完整的预处理过程，通过数据填补、低方差滤波以及高相关滤波，填补了数据缺失部分，并对筛选出的贡献相对较少的维度进行删除，最终得到了一份不包含缺失数据，维度为 208×60 的数据样本。

# 第 3 章　基于拓扑聚类算法的心功能评级

## 3.1　引言

在第二章中，对数据进行了填补删除等预处理操作，本章的任务是实现拓扑聚类算法和 Mapper 算法，并对与算法得到的结果进行分析。

## 3.2　拓扑聚类算法

### 3.2.1　算法介绍

拓扑数据分析（TDA）是针对高维数据分析中聚类分析无法解决的各种问题而产生的。拓扑学关注集合对象的定性性质，因此，它不像纯几何方法那样对坐标或度量的选择那么敏感。对于给定的数据集，TDA 的策略是构建数据的几何表示，并对该表示结果应用拓扑方法，从中可以对数据的形状和连通性进行推断和数据抽象，通过抽象过后得到的拓扑图中染色情况以及点集簇之间的位置关系给出最后的聚类结果。下面一节通过 Mapper 算法实现 TDA 的拓扑数据分析思想并聚类。

### 3.2.2　Mapper 算法实现

首先，使用滤波函数 function，将数据集 Data 映射到实数空间 R 中。常见的滤波函数选择可以是实数值中心性函数（3-1）或者偏心函数（3-2）：

$$\mathrm{F}unction_{r}(x)=\sum_{y\in Data} dis(x,y) \tag{3-1}$$

$$\mathrm{F}unction_{\mathrm{e}}(x)=\max_{y\in Data} dis(x,y) \tag{3-2}$$

然后构造一组对滤波值的覆盖 $\mathrm{C}_i(i\in \mathrm{Q})$，这个覆盖通常采用一组重叠间隔的形式，其中覆盖的恒定长度和覆盖之间的重合百分比由参数 n_intervals 和 overlap_fraction 给出。随后根据得到的对滤波数据的覆盖结果，通过逆映射 $\mathrm{Function}^{-1}$ 找到每一个覆盖所对应的进行滤波之前的数据原本值，对每一个覆盖内部采取聚类算法（本课题中使用 DBSCAN）进行聚类，聚类算法得出的每一个聚簇即为一个拓扑图中的点。最后将相邻的覆盖之间通过聚类算法抽象得到的数据点通过边链接，得到最终的拓扑聚类算法的拓扑图，并根据图中的数据间的连通性得到聚类结果。

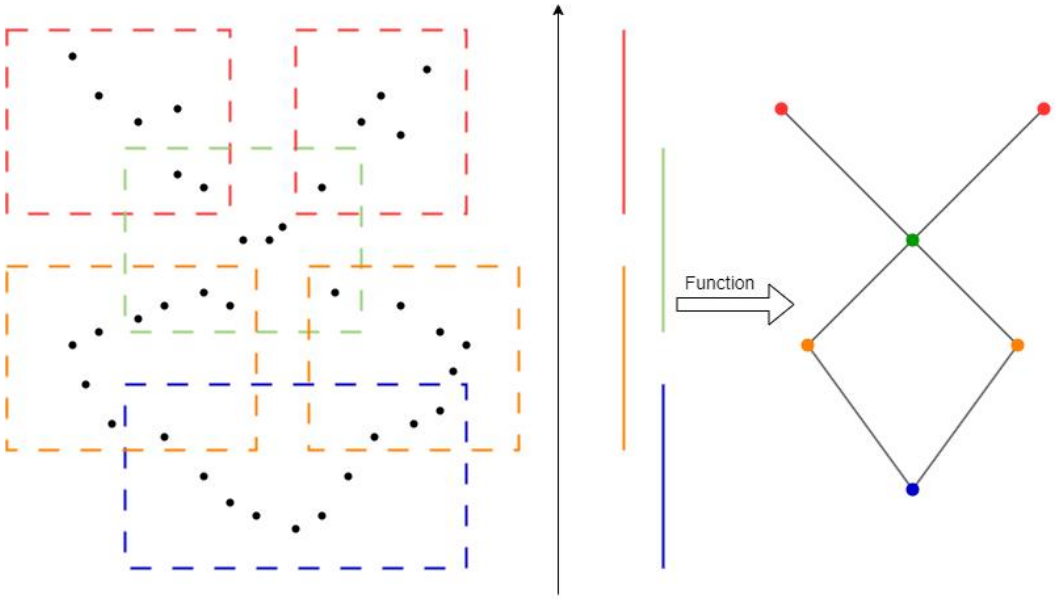


图 3-1 Mapper 算法图示

## 3.3 具体实现以及聚类结果

首先定义 Mapper 算法中滤波函数的距离度量，公式如下所示：

$$dis(x,y)=1-Pearson(x,y) \tag{3-3}$$

其中 Pearson 为皮尔逊相关系数，公式由上一章节中给出。接下来需要对 Mapper 算法中参数 n_components、n_intervals 和 overlap_fraction 进行遍历比较，其中 n_components 为输入数据维度，可以通过主成分分析 PCA 算法[16]进一步降维到选定的值；而恒定覆盖长度 n_intervals 和覆盖重合度 overlap_fraction 则通过遍历参数组合列表来进行实验。实验过程中得到的较好的效果如下：

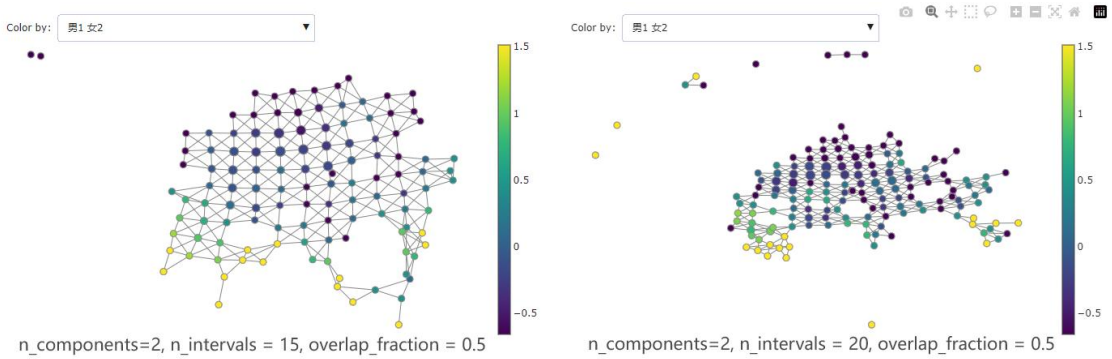


图 3-2 较好聚类结果

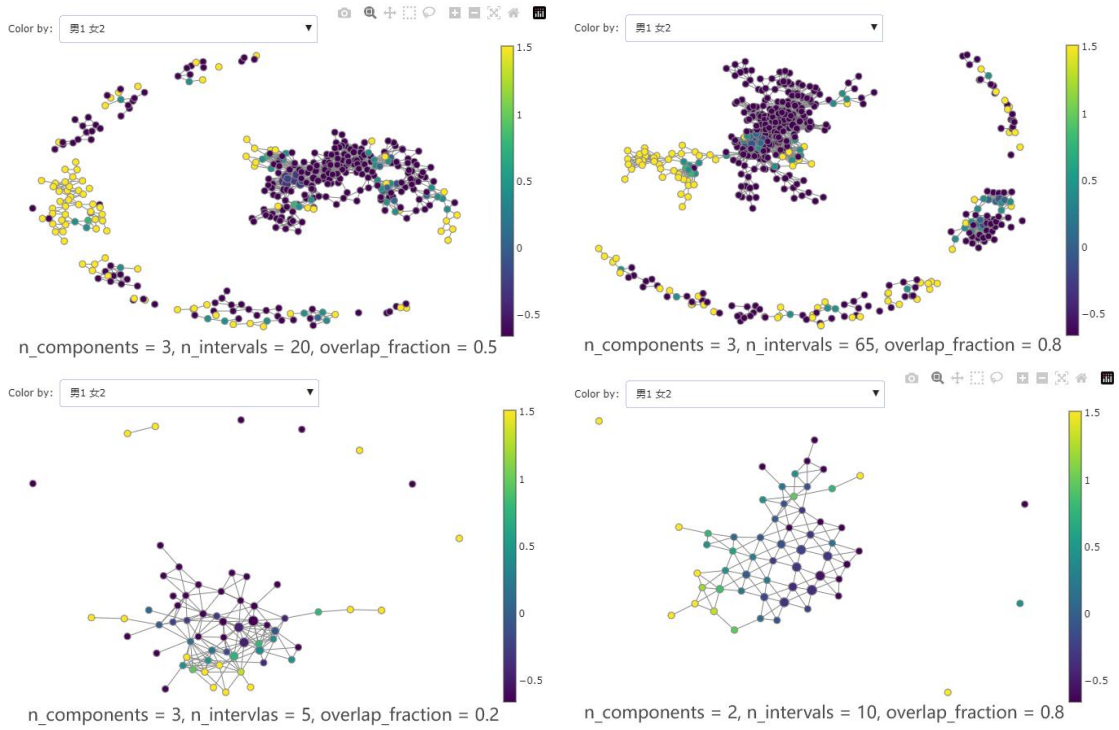


图 3-3 较好聚类结果（续）

通过对实验结果的观察，发现效果最好的参数组合为[n_components = 2, n_intervals = 15, overlap_fraction = 0.5]。根据拓扑图的结构以及染色情况可以将208 个样本聚类为三类，其中每一个点代表着一个或几个样本数据，具体聚类情况如下图所示：

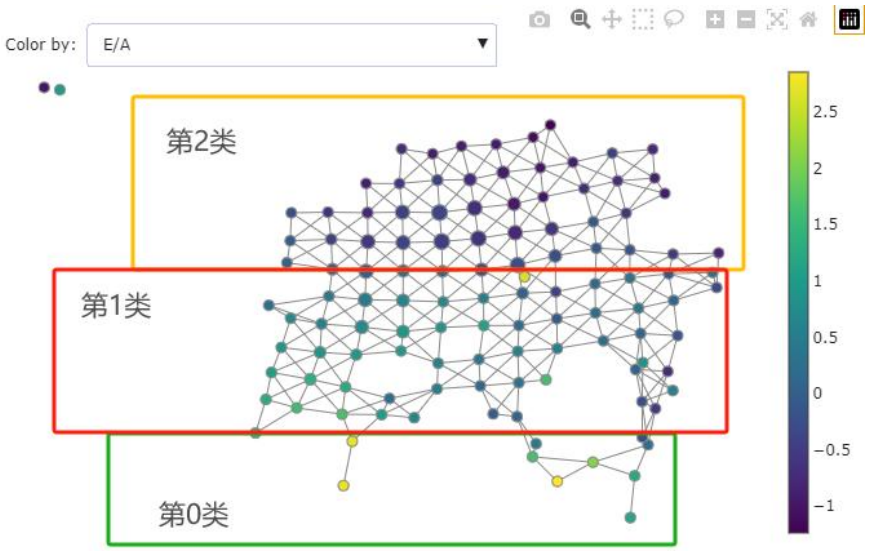


图 3-4 聚类结果划分

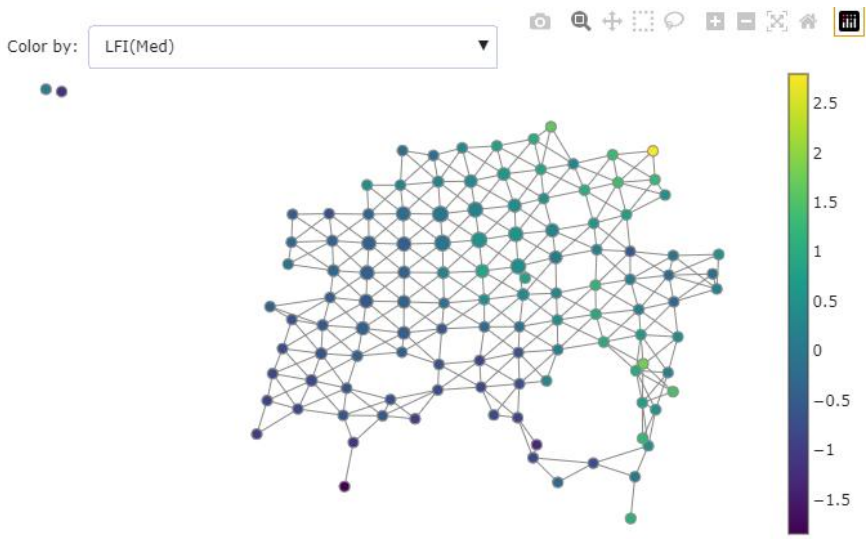


图 3-5 聚类结果的另一种染色

提取聚类标签，对应于原始数据，我们可以得到每个聚类类别的数字特征信息进行简要的分析。通过对聚类类别数组特征信息的观察，结合指标的医学含义以及作用，发现 VFM-相对压中射血收缩期的 IVPG 值在类别 0 和类别 1 中，以及 VFM-能量损耗中心房收缩期 SUM 在类别 1 和类别 2 中有较为明显差别，如下表所示：

表 3-1 区别较大指标

| 维度名称 | 第 0 类 | 第 1 类 | 第 2 类 |
|---|---|---|---|
| 射血收缩期 IVPG | -0.054±0.018 | -0.040±0.022 | -0.044±0.021 |
| 心房收缩期 SUM | 0.056±0.061 | 0.052±0.040 | 0.043±0.042 |

其中这两个指标，射血收缩期 IVPG 来自于 VFM-相对压，心房收缩期 SUM 来自于 VFM-能量损失，通过对这两个生理指标的了解我们得知，VFM-相对压产生于心脏舒张早期，从心脏基底部分到心尖部分出现的负的舒张期抽吸，因此数值越小代表心脏功能越好；而 VFM-能量损失则代表着由于血液粘滞性，在湍流出现的地方产生的摩擦生热，因此能量损失越小说明心脏功能越好。由此可见，第 0 类的患者心脏功能是最强的，第 1 类则次之，第 2 类患者心脏功能则最差，需要进行调理。

至此，拓扑聚类算法结束，聚类结果表明该算法能够将高维数据样本分类并得到较为合理的聚类结果。

## 3.4 本章小结

在这一章中我们对拓扑聚类算法进行了介绍，并通过 Mapper 算法实现了拓扑聚类分析，通过对高维开方体的维数、恒定覆盖长度以及两个相邻覆盖之间重合比率进行选择和组合，通过不同的参数组合构成不同的拓扑聚类算法模型的结构，并将经过处理的 208 个样本投入模型中进行训练，最终确定了最佳参数组合为[n_components = 2, n_intervals = 15, overlap_fraction = 0.5]。模型将 208 个数据分为了三类，计算了每个类别的数字特征，结合医学指标意义进行了分析，说明了拓扑聚类算法能够将数据进行合理聚类。

# 第 4 章　基于自编码器深度聚类算法的心功能评级

## 4.1　引言

在第三章中实现了拓扑聚类分析并且得到了合理的聚类结果，在本章节中，将实现构建全连接自编码器和卷积自编码器，并使用两种不同网络架构的自编码器进行深度神经网络聚类算法，对预处理过后的数据进行聚类，并分析聚类结果。

## 4.2　算法原理

### 4.2.1　全连接自编码器

自编码器算法的主要目的是学习输入数据的信息性，这个信息性表示通过对数据数据的学习，尽可能多的学习到数据维度之间的关系，对数据维度进行降维、融合，尽可能地使用神经网络吸收数据的各种特性。自编码器的典型架构如下图所示，主要组件有编码器、潜在特征表示器和解码器：

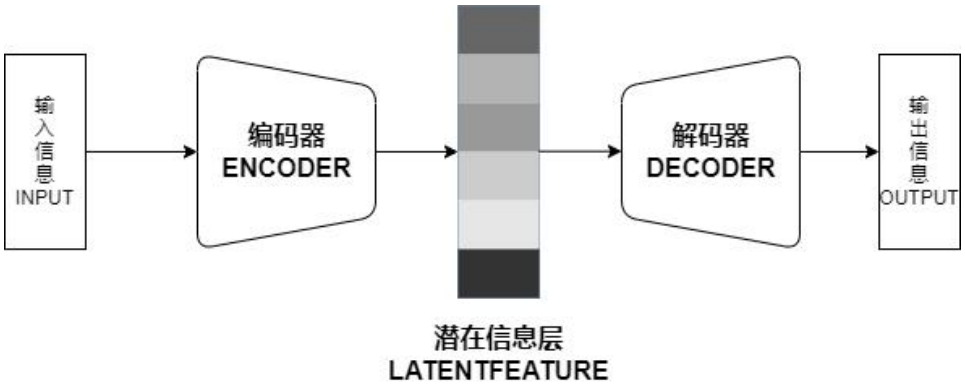


图 4-1 自编码器典型架构

其中编码器和解码器一般都以神经网络的形式架构。自编码器的核心思想是：将输入数据投入到神经网络中，经过对称的神经网络对数据进行编码和解码，最终能够得到一份输出，自编码器神经网络的训练目标正是让输入数据和输出数据尽可能的相似。在训练结束之后，提取自编码器中的潜在信息层作为输入数据的信息代表，潜在信息层的维度更小，聚合性更强，每一个维度都能代表原有数据的多个维度，以达到学习数据信息性的目标。

在本课题中，首先实现了一个全连接层结构的自编码器。全连接神经网络

也被称为多层感知机，其网络结构经典架构如图 4 - 2 所示。

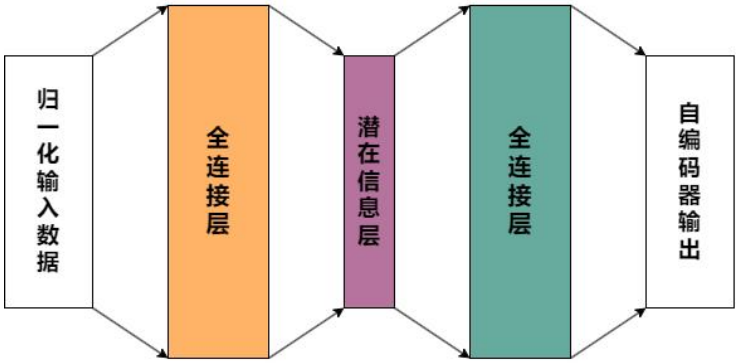


图 4-2 全连接神经网络

全连接自编码器[17]也被称为前馈自编码器，他是一个对称的结构，以潜在信息层的中心为轴，前一部分的神经元数量逐渐减少，后一部分的神经元数量逐渐增加，呈对称的形式，具有奇数层数。对于输出层的激活函数，本课题选取 ReLU 函数，其公式如下：

$$\mathrm{ReLU}(x) = \max(0, x) \tag{4-1}$$

在进行自编码器的训练之前，本课题会将输入数据进行 min - max 归一化，使得输入数据都分布在闭区间[ 0 , 1 ]中，输入观测值全为正值，因此选择 ReLU 函数作为输出层的激活函数。全连接自编码器的损失为输入观测数据和输出重构数据的差别大小，可以将这一问题看做一个广义回归问题，于是本课题选择使用均方损失（MSE）作为全连接自编码器的损失函数，其公式定义定义如下所示：

$$L_{MSE} = MSE = \frac{1}{M}\sum_{i=1}^{M} |x_i - \tilde{x}_i|^2 \tag{4-2}$$

其中 M 表示输入数据的维度，即数组长度；$x_i$ 表示输入数据的某一维度数值，$\tilde{x}_i$ 表示对应维度信息值的预测值。

构建完成全连接层网络的结构之后，通过损失函数和网络结构使用反向传播算法 BP 进行网络中权重值的更新。反向传播算法针对神经网络中的神经元单位进行调整，通过反复更改网络中连接权重的方式，计算损失，优化损失函数输出，以最小化网络的实际输出向量与期望输出向量之间的差异。在将输入数据作为 INPUT 放入神经网络中之后，通过前馈神经网络正向传播的一系列计算，数据通过编码器部分、潜在信息层部分以及解码器部分之后，产生了一定的误差，通过损失函数 $L_{MSE}$ 进行计算。而损失函数中包含着紧邻输出层 OUTPUT 前一层的连接权重参数，对损失函数进行求导，计算方向导数方向的

反向，即可得到前一层的连接权重参数需要进行更新的方向。并通过设置学习率，也就是更新的步长，即可得到前一层的参数需要更新得到的具体数值。通过使用函数求导的链式求导法则，可以从输出层开始进行反向传播计算，如此便可进行整个全连接自编码器网络中连接权重参数的更新，反向传播算法总结如下：

**算法 2** 反向传播算法

**输入：**网络结构权重序列 H，输入数据 X

**输出：**更新后权重序列 H’

**1：**设置损失函数 Loss [ x , H ( x ) ]和训练停止阈值 Loss Limit

**2：while** Loss（x） > Loss Limit，**do：**

**3：** 通过函数求导的链式求导法则计算每一层方向导数

**4：** 向方向导数的反方向进行权重参数更新，更新为 H’

**5：** 更新损失 Loss [ x , H ’ ( x ) ]

**6：end while**

**7：return** H’

通过不断地进行训练，使得损失函数逐渐降低，降低到预先设置的 Loss 停止值则说明模型已经训练到满意的程度，停止训练，提取潜在信息层的数据矩阵，供后面小节深度聚类算法使用。至此，全连接自编码的架构和训练内容结束，下一小节进行卷积自编码器的架构和训练。

### 4.2.2 卷积自编码器

本课题的数据量较小，通过采取马尔可夫变迁场[18]进行数据升维操作，并实现一个卷积自编码器，构造卷积核对输入数据进行卷积操作，可以在不歪曲数据自身特征的情况下更大程度的利用输入数据的维度信息，提升聚类效果以及自编码器的网络质量。

马尔可夫变迁场是一种将顺序序列数据转换成二维图像的算法，它利用了马尔可夫链模型，通过计算数据间的转移概率，构造马尔科夫转移特征向量，来达到使一维数据变为二维数据的效果。马尔可夫变迁场算法的整体流程如下页伪代码算法 3 所示。

对于自编码器的对称结构，本小节构造一个全卷积神经网络[19]自编码器，对于任意输入尺寸的数据，经过推理和学习之后能够直接得到相应尺寸的输出，再通过均方损失函数对网络连接权重参数进行优化。本小节构造了卷积神经网

络和反卷积网络。

**算法 3** 马尔可夫变迁场

**输入：**一维数据 X

**输出：**升维后的数据矩阵 Matrix

**1：**设置子序列长度 L

**2：**将输入的一维数据 X 分割成长度为 L 的彼此不重叠子序列

**3：For** 每一个子序列：

**4：** 统计子序列中不同的数据数量，即状态数 K

**5：** 计算相邻数据之间的状态转移概率，状态转移概率矩阵维度为[ K , K ]

**6：** 将状态转移矩阵中的每一个元素作为特征向量的一个元素，得到该子序列的马尔科夫转移特征向量

**7：**拼接所有子序列的马尔科夫转移特征向量，构成 Matrix

**8：return** Matrix

卷积神经网络层级结构主要有五个部分组成，包括数据输入层、卷积计算层、ReLU 激励层、展平层和全连接层，具体结构以及各层数据张量大小将在本章下一小节——算法实现中给出。

首先是数据输入层。通过马尔可夫变迁场的数据升维构造，我们已经得到了一系列在不歪曲数据本身特征情况下，进行合理升维的二维数据。由于数据的取值范围上下界的不统一，直接作为数据输入进行训练会影响模型的准确性和训练效果，因此在数据输入之前要进行数据归一化，同样采用 min - max 归一化方法，随后将归一化后数据投入卷积计算层进行计算。在进行卷积计算之前，需要进行可选择的填充操作，即如果相邻两个层之间的数据维度不匹配导致无法进行卷积计算时，通过零填充 zero padding 方法对数据进行维度填充，以保证数据维度符合相邻两个层之间的计算要求。通过相邻两个卷积计算层的输出数据维度，即可计算卷积感受野以及卷积步长 stride，将感受野中包含的数据投入卷积权重参数，进行计算得出感受野中数据经过卷积计算层之后的结果。具体过程如图 4 - 3 所示。

卷积层的输出结果需要进行非线性映射后再进行传播，较为常用的即为 ReLU 激活函数，函数公式在上一小节中给出。在将数据投入到全连接层之前，需要将卷积层数据进行张量扁平化操作，展平层的工作目标就是将二维卷积数据转化为一维数据提供给全连接层的计算。全连接层即位卷积操作的终点，通过激活函数的计算过后得到的数据即为整个卷积自编码器的编码器部分和潜在信息层部分的训练构建，全连接层的输出作为潜在信息层的数据被提取保存下

来，提供给下一小节中深度聚类算法使用。在卷积操作完成之后，按照卷积自编码器的架构，需要以全连接层的输出数据为输入，进行反卷积操作，构建自编码器解码器部分的输出。

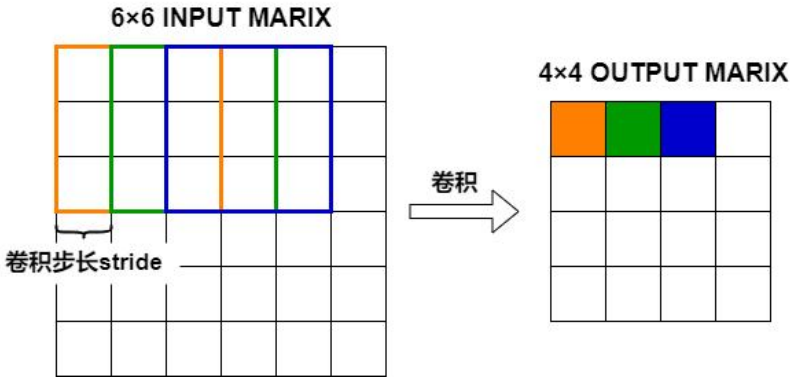


图 4-3 卷积计算层操作图示

反卷积操作即为正向卷积特征提取操作的逆过程，同样通过计算相邻两个反卷积计算层的输出数据维度，得到感受野和步长 stride，将前一反卷积层的数据矩阵扩大，这一操作被称为上采样过程。在经过了一系列的上采样过程，最终能够由对称的卷积自编码器网络结构得到一个和输入数据维度相同的数据输出。最后计算输入输出二维数据的均方损失，通过损失函数的方向导数反向以及反向传播算法更新网络结构。至此，卷积自编码器的架构与训练工作结束，下一小节进行深度聚类算法的架构。

### 4.2.3 深度聚类算法

考虑将 n 个点聚类成 k 个簇，相比于直接在高维度样本空间 X 中直接进行聚类，深度聚类算法[20]首先用一个非线性映射

$$Function_{\theta}: X \rightarrow Z \qquad (4\text{-}1)$$

将高维度空间 X 映射到低纬度的空间 Z 中，其中 θ 表示可学习参数，Z 为潜在信息空间，并选择使用神经网络来进行函数特征学习以及参数化 $Function_{\theta}$。

深度聚类算法通过同时学习潜在特征空间 Z 中的 k 个聚类中心和神经网络的参数 $Function_{\theta}$来聚类数据。算法有两个阶段：

1）用深度自编码器进行参数初始化；

2）进行参数优化，通过迭代计算辅助目标分布和最小化 KL 散度。其中 KL 散度计算由公式（4 - 2）至（4 - 4）给出，其中 $p(x_i)$表示变量 X 取某一值的概率，P 和 Q 表示两个不同的离散随机变量的分布，CrossEntropy 表示两个分布之间的交叉熵损失。

$$H(X)=-\sum_{i=1}^{k}p(x_i)\log p(x_i) \tag{4-2}$$

$$\text{Cross}Entropy(P,Q)=\sum_{i=1}^{k}P(x_i)\log\frac{1}{Q(x_i)} \tag{4-3}$$

$$KL(P\|Q)=-H(P)+CrossEntropy(P,Q) \tag{4-4}$$

深度聚类算法参数优化方法：对于给定非线性映射 $\text{Function}_\theta$ 的初始估计和初始聚类中心 $\{u_j\}_{j=1}^{k}$，通过使用在两个步骤之间交替的无监督算法来改进聚类，算法梗概总结如下：

**算法 4** 改进聚类算法

**输入：**非线性映射 $\text{Function}_\theta$ 和初始聚类中心 $\{u_j\}_{j=1}^{k}$

**输出：**满足置信度要求的算法参数

**1：**设置置信度要求 tolerance

**2：while** step_tolerance > tolerance，**do：**

**3：** **step1：**计算样本点和聚类中心之间的软分布，软分布由 Student t 分布给出，表示样本点属于该簇的概率

**4：** **step2：**通过使用构造目标分布从当前置信度分配中学习来改进聚类中心

**5：** 更新置信度 step_tolerance

**6：end while**

**7：return** $\text{Function}_\theta$，$\{u_j\}_{j=1}^{k}$

其中第一步为计算软分配。以 t 分布作为内核来测量样本点 $z_i$ 和聚类中心 $u_j$ 之间的相似性，公式如下：

$$q_{ij}=\frac{(1-\|z_i-u_j\|^2\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j'}(1-\|z_i-u_j\|^2\alpha)^{-\frac{\alpha+1}{2}}} \tag{4-5}$$

其中 $z_i=\text{Function}_\theta(x_i)$，X 和 Z 即为上述提到的高维数据空间和低维数据空间以及非线性映射关系；α 为 t 分布的自由度，决定了 t 分布的形状以及影响 t 分布的置信区间；$q_{ij}$ 可以解释为第 i 个样本被预测为第 j 个簇类的概率，即软分配。

通过第一步，得到了预测分布 Q 以及每个样本点的软分配，接下来需要构造出一个目标分布 P，并让预测分布通过学习训练尽可能地逼近目标分布，降低两个分布之间的 KL 散度。

理想的构造目标分布 P 应该具备以下属性：

1）提高聚类纯度

2）加强高置信度软分配

3）规范化每个聚类中心的损失贡献，以防大团簇聚类扭曲隐藏特征空间

为了得到这样一个构造目标函数 P，首先 $q_i$ 提高到二次幂，然后按照每个簇的频率来进行归一化计算 $p_{ij}$，公式如下：

$$p_{ij} = \frac{q_{ij}^2 / Function_j}{\sum_{j'} q_{ij'}^2 / Function_{j'}} \tag{4-6}$$

其中 $Function_j$ 为 $q_{ij}$ 对 i 求和。这样的训练策略实质上就是自我训练的一种形式，通过对高置信度预测的学习来改善低置信度预测，在每一次循环训练的过程中，算法通过学习高置信度预测来改进初始估计，反过来又能作用与低置信度预测，帮助改进预测结果。

参数优化采取了带动量的随机梯度下降[21]，即 SGD 算法对聚类中心和深度神经网络的参数 $Function_\theta$ 进行联合优化。梯度下降算法通过计算损失函数方向导数最大值，即梯度方向，随后沿着梯度方向的反方向进行权重的更新，能够更加快速的收敛到最优解的位置。随机梯度下降算法即从样本中随机抽取一组，训练过后按照梯度更新一次，随后重复迭代这个操作，最终使得损失值，即 KL 散度达到预先设计的分布相似度要求，停止训练。SGD 算法流程如下：

**算法 5** 随机梯度下降算法

**输入：**损失函数停止界限 Loss Limit，输入数据 X

**输出：**网络参数

**1：**设置损失函数 $J(\theta) = (1/2) \times SUM[h_\theta(x) - y]^2$

**2：while** Current Loss > Loss Limit，**do：**

**3：** 随机取出输入数据中的某一组数据，计算损失函数 $J(\theta)$，更新 Current Loss = $J(\theta)$

**4：** 计算梯度方向，向着梯度方向的反方向进行权重更新

**6：end while**

**7：return** $h_\theta$

损失函数 L 对于每个数据点 $z_i$ 和每个聚类中心 $u_j$ 的梯度计算如公式（4 - 7）和（4 - 8）所示。然后将梯度公式（4 - 7）的计算结果传递给深度神经网络，并在标准反向传播中计算深度神经网络的参数。当两次联系迭代训练之间更改聚类分配的点少于 tolerance 时，表明训练结束，达到目标并停止训练。至此，

深度聚类算法的架构和训练工作内容结束。

$$\frac{\partial \mathrm{L}}{\partial z_i}=\frac{\alpha+1}{\alpha}\sum_j(1+\frac{\| z_i-u_j \|^2}{\alpha})^{-1}\times(p_{ij}-q_{ij})(z_i-u_j) \tag{4-7}$$

$$\frac{\partial \mathrm{L}}{\partial u_i}=\frac{\alpha+1}{\alpha}\sum_i(1+\frac{\| z_i-u_j \|^2}{\alpha})^{-1}\times(p_{ij}-q_{ij})(z_i-u_j) \tag{4-8}$$

## 4.4 算法实现

首先实现全连接自编码器架构。在第二章中，经过预处理过程得到了数据维度为[ 208 , 60 ]的完整数据，在进行自编码器的训练之前，按照上一小节中的架构流程，需要进行数据归一化，归一化后的部分数据如下表所示：

表 4-1 归一化后数据样本（部分）

| 样本 | 维度名称 | | | | | | |
|---|---|---|---|---|---|---|---|
| | 男 1 女 2 | 年龄 | 身高 | 总蛋白 | 直接胆红素 | …… | 心房收缩期 |
| 1 | 0.00 | 0.59 | 0.57 | 0.81 | 0.05 | …… | 0.11 |
| 2 | 0.00 | 0.63 | 0.58 | 0.79 | 0.14 | …… | 0.15 |
| 3 | 0.00 | 0.61 | 0.40 | 0.71 | 0.09 | …… | 0.12 |
| | | | | …… | | | |
| 204 | 0.00 | 0.31 | 0.63 | 0.85 | 0.55 | …… | 0.16 |
| 205 | 0.00 | 0.55 | 0.61 | 0.90 | 0.20 | …… | 0.36 |
| 206 | 0.00 | 0.82 | 0.42 | 0.92 | 0.02 | …… | 0.27 |

接下来进行自编码器的预训练过程。首先是超参数的设置，参数 dim 表示编码器中每层的单元数量列表，编码器和解码器对称，计算出全连接自编码器的层数 n_stacks = len ( dim ) - 1。预训练网络权重参数优化器选择随机梯度下降 SGD 优化器，预训练轮数 pretrain_epochs 设置为 500，数据批量大小 batch_size 设置为 64，即每轮训练投喂 64 个随机样本数据进自编码器全连接神经网络，损失函数选择均方损失 MSE，超参数设置完成之后开始训练并保存网络中的权重参数。全连接自编码器的网络结构图在上一章中已经给出，如图（4 - 2）所示。

自编码器训练结束之后，进行深度聚类层的构建。第一步，使用 KMeans[22] 方法初始化聚类中心，通过第三章实现的 TDA 拓扑聚类算法，得知将聚类中心数量 n_clusters 初始化为 3 是较为合理的选择，同时也考虑将聚类中心数量提高到 4 和 5 来进行对比，最终选择出最为合适的聚类中心数量。第二步，按

照上一小节中的算法，构建出目标分布，设置前后两次深度聚类样本点分布差异值阈值 tolerance，当前后两次迭代的结果差异值小于 tolerance，则说明训练完成，并保存深度神经网络的权重参数。经过自编码器的训练和深度聚类算法的聚类，最终得到聚类结果 y_pred。

将聚类结果和第三章中实现的 TDA 拓扑聚类算法得到结果进行对比，可以得到两种聚类结果标签的混淆矩阵，如下图所示：

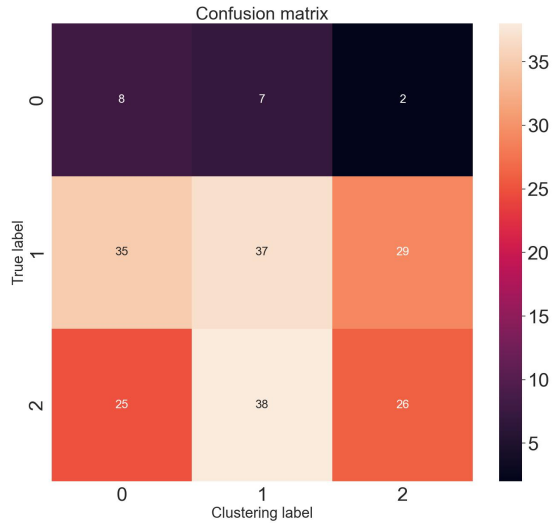


图 4-4 两种聚类算法结果混淆矩阵

根据混淆矩阵结果，发现两种算法得到的结果具有较大的差异，对于聚类算法得到结果的合理性分析将在下一小节中进行，不同聚类模型优劣对比将在下一章中详细阐述。

接下来实现卷积自编码器架构。根据反卷积操作的不同，本课题实现了两种卷积网络的构造并分别进行实验，对比分析并选取效果较好的卷积网络作为本小节的卷积网络构造。两种卷积神经网络的主要区别在于上采样方法不同，分为了 Conv2DTranspose 上采样和 UpSampling2D 上采样。其中 Conv2DTrans-pose 与上一小节中算法原理介绍中的上采样操作做法一直，而 UpSampling2D 的反卷积操作有所不同，它可以看作是池化操作的反向操作，采用了最近邻插值算法来对输入数据进行放大，以实现反卷积操作的效果。对于两个反卷积层，前一层的数据矩阵尺寸必然小于后一层，根据反池化的思想，将后一层的数据矩阵缩小对照在前一层的数据矩阵中，通过计算后一层中每一个点在前一层中最相邻的位置来确定反池化操作的结果，计算公式如下：

$$nextX = previousX \times nextWidth / previousWidth \quad (4\text{-}9)$$

$$nextY = previousY \times nextHeight / nextHeight \quad (4\text{-}10)$$

其中( nextX , nextY )为后一层数据矩阵中某一个数据对应的矩阵坐标，

( previousX , previousY )为前一层数据矩阵中某一个数据对应的矩阵坐标，Width 和 Height 分别代表了对应数据矩阵的两个维度，即长和宽。最近邻插值算法流程总结和图示如下：

**算法 6** 最近邻插值算法

**输入：**前一层数据矩阵 $Matrix_{pre}$

**输出：**后一层数据矩阵 $Matrix_{next}$

**1：For** ( nextX , nextY ) **in** $Matrix_{next}$，**do：**

**2：** 计算与矩阵 $Matrix_{pre}$ 中最相邻的位置( preX , preY )

**3：** 将 $Matrix_{next}$ [ nextX , nextY ]的值置为 $Matrix_{pre}$ [ preX , preY ]

**4：return** $Matrix_{next}$

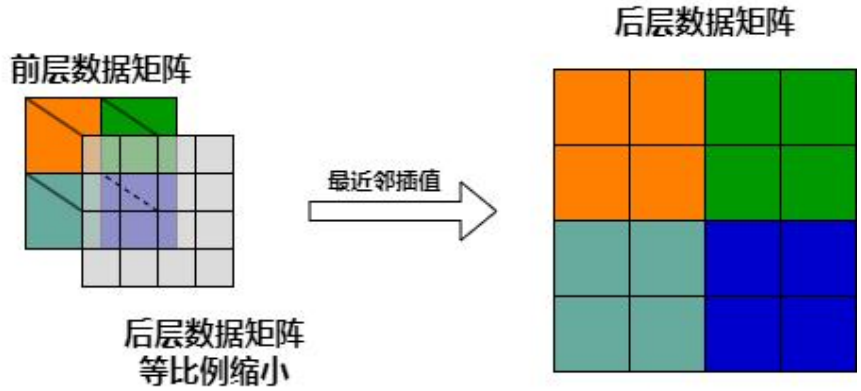


图 4-5 最近邻插值算法图示

两种卷积神经网络的具体结构以及神经元数量如图（4 - 6）和图（4 - 7）所示。最后设置训练轮数 pretrain_epochs = 100，批量大小 batch_size = 64 对两种卷积神经网络进行训练，得到聚类结果的预测标签 y_pred，聚类结果的对比分析将在下一小节中进行。至此，自编码器训练任务的全部工作结束，下一小节中将对聚类的得到的结果进行分析，选择效果最好的数据升维方式和编码器网络结构组合。

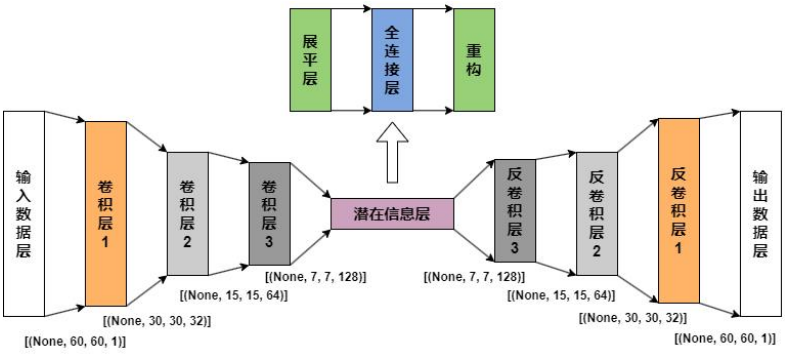


图 4-6 Conv2DTranspose 方法卷积神经网络

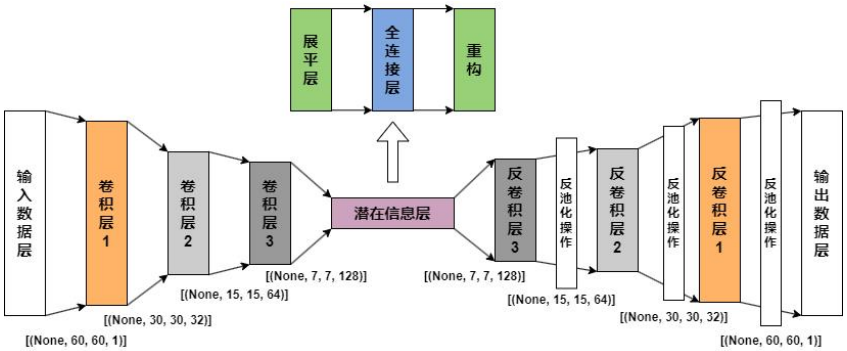


图 4-7 UpSampling2D 方法卷积神经网络

## 4.5 聚类结果分析

在进行结果分析之前，为了对比不同模型之间聚类结果的好坏，选择使用双样本 K - S 检验来判断两个类别同一维度是否具有显著差异。双样本 K - S 检验公式如下：

$$Distance_{n,m} = \max\left[ CDF\left(X_1 \text{ distribution}\right) - CDF\left(X_2 \text{ distribution}\right)\right] \quad (4\text{-}11)$$

其中 n 和 m 表示两个维度的数据个数，CDF 表示两个样本的分布函数。经过推导，p - value 值的公式如下所示：

$$\text{p - value} = 2e^{-2(Distance_{n,m})^2 \frac{nm}{n+m}} \quad (4\text{-}12)$$

p - value 值可以解释为：值越小，越有理由认为两个数据样本来自于不同的分布，即两个维度信息的数据区分度越高，聚类效果越好。本课题中认为，p - value 值小于 0.01，则十分有理由认为两个数据样本有极大可能性来自于不同的数据分布，即不同等级的心脏功能分级；p - value 值介于 0.05 和 0.1 之间，则有一定理由认为两个数据样本较有可能来自于不同的数据样本。

### 4.5.1 聚类中心个数分析

根据第三章中的 TDA 拓扑聚类算法得到的结果，进行自编码器训练时首先将聚类中心个数 n_clusters 设置为了 3 进行试验，不失一般性，对于更高的聚类中心个数也进行了独立实验进行讨论。首先是将 n_clusters 设置为 4，分别对四个样本聚类两两之间进行 K - S 检验，统计两两聚类之间的 p - value 值。经过统计发现第 0 类样本在和其余三类样本进行 K - S 检验之后，没有任何一个维度的 p - value 值在进行三次检验时同时小于 0.1，在进行第 3 类样本的检

验时也有同样的结果。这个结果说明在当前的聚类结果中，从第 0 类或者第 3 类样本中抽出一个样本个体，没有充分理由认为能够判断这个样本个体来自于哪一个聚类类别。进一步对每两个聚类样本都进行 K - S 检验，发现第 1 类样本、第 2 类样本和第 3 类样本之间的 p - value 值较为理想，如下表所示：

表 4-2 第 1 类样本和第 3 类样本 K - S 检验结果

| 维度名称 | P - value 值 |
|---|---|
| 二尖瓣 E 峰减速时间 DT cm/s | $p < 0.01$ |
| VFM-能量损耗等容收缩期 AVE 重点考虑提标面积标化最大值 | $p < 0.01$ |
| VFM-能量损耗收缩射血期 AVE 重点考虑提标面积标化平均值 | $p < 0.01$ |
| VFM-能量损耗收缩射血期 AVE 重点考虑提标面积标化最大值 | $p < 0.01$ |
| VFM-能量损耗心房收缩期 SUM 最大值 | $0.01 < p < 0.05$ |
| VFM-能量损耗等容收缩期 AVE 重点考虑提标面积标化中位数 | $0.01 < p < 0.05$ |
| VFM-能量损耗收缩射血期 SUM 最大值 | $0.01 < p < 0.05$ |
| VFM-能量损耗收缩射血期 AVE 重点考虑提标面积标化中位数 | $0.01 < p < 0.05$ |
| VFM-能量损耗心房收缩期 AVE 重点考虑提标面积标化最大值 | $0.01 < p < 0.05$ |
| INR 凝血酶原时间 | $0.05 < p < 0.1$ |
| 腹部门静脉内径 正常 0 增宽 1 | $0.05 < p < 0.1$ |
| 常规心脏超声左室舒张指标 A | $0.05 < p < 0.1$ |
| VFM-能量损耗收缩射血期 SUM 中位数 | $0.05 < p < 0.1$ |

表 4-3 第 2 类样本和第 3 类样本 K - S 检验结果

| 维度名称 | P - value 值 |
|---|---|
| VFM-能量损耗收缩射血期 SUM 最大值 | $0.01 < p < 0.05$ |
| VFM-能量损耗收缩射血期 AVE 重点考虑提标面积标化最大值 | $0.05 < p < 0.1$ |
| VFM-相对压舒张缓慢充盈期 IVPG | $0.05 < p < 0.1$ |
| 肝脏联合弹性指标 FIndex（Med） | $0.05 < p < 0.1$ |

由此结果认为，将聚类中心数量提高到 4 个，会出现某些聚类类别之间分界线模糊的问题，判断为将本不需要归为两类的样本个体划归到不同类别中导致。随后继续将聚类中心数量提高到 5 个进行对比实验，多个聚类结果两两之间的 K - S 检验结果没有同时达到 p - value 值小于 0.1 的信息维度，为不必要的聚类区分。综上所述，将聚类中心个数 n_clusters 设置为 3 是较为合理的选择，最终确定聚为 3 类。

### 4.5.2 卷积神经网络结构分析

经过上一小节的训练，最终得到了三个神经网络的聚类结果，下面进行不同网络结构之间的结果对比与分析。

首先是全连接层自编码器的类间双样本 K - S 检验。对于 p - value 值小于 0.01 的数据维度称之为两个聚类样本的有效区分，三类样本两两之间的有效区分如下表所示：

表 4-4 三类样本两两之间的有效区分维度

| 第 0 类和第 1 类有效区分维度 | 第 0 类和第 2 类有效区分维度 | 第 1 类和第 2 类有效区分维度 |
|---|---|---|
| 肝脏联合弹性指标 FIndex | 肝脏联合弹性指标 ATT | 肝脏联合弹性指标 ATT |
| VFM 等容 AVE 中位数 | VFM 缓慢充盈期 IVPG | VFM 等容 AVE 中位数 |
| VFM 心房 AVE 最大值 | VFM 充盈期 AVE 中位数 | VFM 心房 AVE 最大值 |
| VFM 心房 AVE 中位数 | | VFM 心房 AVE 中位数 |
| 体重（kg） | | VFM 心房 SUM 最大值 |
| INR 凝血酶原时间 | | |
| VFM 等容收缩期 IVPG | | |
| VFM 等容射血期 IVPD | | |
| VFM 缓慢充盈期 IVPG | | |

通过数据表格发现经过全连接层自编码器的聚类，两两样本类别之间均有能够明显区分两个聚类类别的数据维度。但是综合来看，没有任何一个数据维度在三个样本之间的所有彼此 K - S 检验结果 p - value 均小于 0.1，因此认为全连接层自编码器能够对患者的心脏功能进行有效评级，但是聚类效果在某些维度上表现一般。接下来观察卷积自编码器的聚类结果。

在上一小节中根据解码器部分上采样操作不同而构建了两种不同的卷积神经网络结构：Conv2DTranspose 和 UpSampling2D。首先对 Conv2DTranspose 网络结构的聚类结果进行分析。通过实验，发现此网络结构得到的聚类结果两两类别之间有较多维度的双样本 K - S 检验 p - value 值小于 0.01，这意味着这些维度在不同的聚类类别之间具有显著的差异。我们将这些维度描述为属于强有效区分的范畴，可以显著的体现出不同聚类结果之间的类间区分，具体结果如下表所示（其中加粗指标为三类样本共同强有效区分维度）：

表 4-5 三类样本两两之间的强有效区分维度

| 第 0 类和第 1 类强有效区分维度名称 | 第 0 类和第 2 类强有效区分维度名称 | 第 1 类和第 2 类强有效区分维度名称 |
|---|---|---|
| GTT 谷氨酰转肽酶 | 肝脏联合弹性指标 ATT | 超声肝病分类 |
| **INR 凝血酶原时间** | **INR 凝血酶原时间** | **INR 凝血酶原时间** |
| 超声肝病分类 | MELD 评分 | GTT 谷氨酰转肽酶 |
| **VFM 射血期 IVPD** | **VFM 射血期 IVPD** | **VFM 射血期 IVPD** |
| **肝脏弹性联合指标 LFI** | **肝脏弹性联合指标 LFI** | **肝脏弹性联合指标 LFI** |
| VFM 射血期 AVE 中位数 | | 肝脏分级评分 |
| 左室后壁厚度（舒张期） | | MELD 评分 |
| VFM 充盈 AVE 中位数 | | |
| VFM 心房 AVE 中位数 | | |
| VFM 射血期 AVE 平均值 | | |

通过观察表格数据可以发现，和全连接层自编码器训练得到的聚类结果相比，Conv2DTranspose 卷积自编码器的聚类结果在进行双样本 K - S 检验时，得到的有效区分维度不仅在数量上超过了全连接层自编码的结果，并且卷积自编码器的 p - value 值小于 0.01 的维度数量相对更多；同时，维度 INR 凝血酶原时间、肝脏联合弹性指标 LFI（Med）和 VFM 射血期 IVPD 在三个类别之间的所有检验中 p - value 值均小于 0.1，说明这两个维度的信息可以区分全部的三个聚类类别，在整个数据维度中都属于样本有效区分。由此可以总结出 Conv2DTranspose 卷积自编码器的聚类结果优于全连接层自编码器。

随后用相同的方式进行 UpSampling2D 卷积自编码器的对比实验。发现在给解码器网络的相邻反卷积层之间加入反池化操作之后，自编码器给出的聚类结果 K - S 检验相比于原来的结构更差，表现为 K - S 检验中有效区分唯独数量减少，并且没有数据维度能够对三个聚类类别进行有效区分，在此对于上述结果进行分析。卷积神经网络的聚类结果优于全连接层自编码器，这是由于数据在投喂到卷积神经网络之前进行了数据升维的操作。在数据进行升维的过程中，相当于用数据自身对自身进行一次强化，升维后的结果包含更多的数据信息，因此网络在进行学习的时候能够学习到更多的数据信息性，因此卷积神经网络的结果比全连接层训练结果更优；对于两个卷积神经网络结构，输入数据的维度皆为[ 208 , 60 , 60 , 1 ]，解释为共有 208 个样本，每个样本为 60×60 的矩阵，在使用了反池化操作之后，由于数据矩阵的尺寸较小，反池化得到的结果并不能很好地反映相邻前一层的数据特征，最终使得在进行神经网络反向传播更新

参数时，由于反池化层的影响，导致传递进行过程中出现了偏差，最终导致结果不如不添加反池化操作的神经网络。

通过理论分析和实验结果，在这一小节确定了聚类效果最好的自编码器结构为 Conv2DTranspose 卷积自编码器。

### 4.5.3 数据升维方式分析

上一章节在进行试验的时候使用了两种数据升维方式，分别为马尔可夫变迁场和最近邻插值法进行数据升维。对于两种不同的升维方式进行对比试验，实验过程和上一小节中的步骤相似。通过对结果进行 K - S 检验，两种数据升维方式得到结果的对比如下表所示：

表 4-6 数据升维方式 K - S 检验

| 升维方式 | 三类样本 p - value 值落在取值区间内的维度数量和 | | |
|---|---|---|---|
| | $p < 0.01$ | $0.01 < p < 0.05$ | $0.05 < p < 0.1$ |
| 马尔可夫变迁场 | 2 | 10 | 5 |
| 最近邻插值 | 19 | 8 | 9 |

通过对比发现，反池化思想数据升维方式得到的结果要优于马尔可夫变迁场，对结果不同进行分析：马尔可夫变迁场是在进行时间序列数据时提出的算法，而本课题的数据为患者的体检样本，本身不具有较强的时间性，而且在进行状态定义时，每两个患者样本可以说是完全不同的两个状态，则需要定义 208 中不同的状态。因此在计算马尔科夫转移特征向量时，前后状态的转移矩阵参考价值很小，计算得到的结果矩阵不能很好地反映数据本身的特征。而最近邻插值算法是将样本在横纵两个方向上，通过数据本身的值进行数据升维，和马尔可夫变迁场相比，更能够反应数据本身的特征，因此得到的聚类结果效果较好。

通过 4.5.1 节到 4.5.3 节的实验和结果分析，最终得出聚类效果最好的自编码器结构为初始聚类中心个数 n_clusters 为 3，数据升维方式为线性插值算法，上采样方法为 Conv2DTranspose 的卷积神经网络具有最好的聚类效果。

### 4.5.4 聚类结果的医学合理性分析

通过上述的聚类工作得出的聚类标签，计算每一个类别的均值以及方差。具有代表性的指标数字特征如下表所示：

表 4-7 实验室检查指标

| 维度名称 | 第 0 类 | 第 1 类 | 第 2 类 |
|---|---|---|---|
| 总蛋白 | 65.88±0.44 | 66.20±11.05 | 65.30±11.85 |
| 谷氨酰转肽酶 | 57.84±47.89 | 68.03±120.61 | 78.62±123.76 |
| 中性粒细胞 | 3.16±2.2 | 3.07±1.73 | 3.02±2.22 |
| 血小板 | 192.15±120.49 | 190.12±98.93 | 154.16±93.28 |
| 血肌酐 | 60.23±13.96 | 62.41±14.05 | 64.40±14.61 |
| 凝血酶原时间 | 1.14±0.17 | 1.08±0.20 | 1.24±0.29 |

表 4-8 肝脏检查相关指标

| 维度名称 | 第 0 类 | 第 1 类 | 第 2 类 |
|---|---|---|---|
| LFI（Med） | 2.15±0.87 | 1.86±0.71 | 1.76±0.76 |
| FIndex（Med） | 2.05±0.98 | 1.55±0.94 | 1.79±0.78 |
| ATT（Med） | 0.58±0.10 | 0.58±0.10 | 0.55±0.08 |
| 肝病超声分类 | 1.41±0.80 | 0.91±0.95 | 1.44±0.78 |
| MELD 评分数值 | 3.92±4.00 | 3.39±4.22 | 6.09±4.02 |
| MELD 分级 | 0.01±0.12 | 0.04±0.25 | 0.11±0.36 |
| 实验室 Fib-4 分级 | 0.79±0.87 | 0.62±0.84 | 1.11±0.90 |
| 实验室 APRI 分级 | 0.12±0.32 | 0.12±0.33 | 0.19±0.40 |

表 4-9 常规心脏超声测量相关指标

| 维度名称 | 第 0 类 | 第 1 类 | 第 2 类 |
|---|---|---|---|
| 舒张期间隔厚度 | 8.30±0.92 | 8.02±0.75 | 8.30±1.16 |
| 舒张期左室后壁厚度 | 8.07±8.02 | 7.68±0.77 | 7.937±1.05 |
| 舒张期左室大小 | 47.23±4.30 | 46.87±4.33 | 47.887±4.17 |
| SimEFBi-Plane | 61.00±5.34 | 62.61±5.03 | 62.12±5.71 |
| 左心室收缩 GLS | -22.58±3.19 | -23.10±3.41 | -22.537±4.35 |
| 左心室舒张 E/A | 1.19±0.38 | 1.18±0.35 | 1.16±0.41 |
| 左心室舒张指标 E | 72.08±16.55 | 69.20±15.30 | 69.20±15.91 |
| 左心室舒张指标 A | 64.37±17.57 | 61.78±14.85 | 63.91±18.58 |
| 舒张功能分级 | 0.21±0.41 | 0.11±0.35 | 0.17±0.46 |

对于不同聚类结果的指标数字特征，结合体检指标的医学意义进行医学合

理性分析。实验室检查指标反映了受检患者肝脏功能以及凝血功能等指标，其中谷氨酰转肽酶与氧化应激、炎症以及心血管疾病相关，长存在于心肌组织中。根据谷氨酰转肽酶的变异性与心力衰竭住院风险的相关研究指出：血液中谷氨酰转肽酶含量的升高会增加患者罹患心衰、缺血性心脏病或肝病的风险，由此推断谷氨酰转肽酶含量相对较低的患者心脏功能越强。通过数据，可以得知第0类患者样本的平均血液谷氨酰转肽酶含量为57.84±47.89，第1类患者样本均值含量为68.03±120.61，第2类患者样本均值含量为78.62±123.76，不难发现第0类患者血液谷氨酰转肽酶含量在三个聚类类别中最低，第2类则最高，说明第0类患者的心脏功能最佳，第1类次之，第2类最差。

心脏和肝脏都是人体内重要的血液枢纽器官，在中医中有“心主血，肝藏血[23]”的说法，意为全身血液充盈，肝有所藏，才能发挥其贮藏血液和调节血量的作用，以适应机体活动的需要，心亦有所主，因此肝脏功能的相关指标也应纳入心脏功能评级考虑的范围内。通过数据可以发现，MELD终末期肝病模型在三个类别中有较为明显的差距。MELD模型反映了肝脏功能的代谢能力，能够作为肝功能的代表数据，其评分值为0代表低危，1代表中危，2代表高危，第0类患者样本的评分均值为0.01，第1类患者样本的评分均值为0.04，第2类患者样本的评分均值为0.11。可以发现，第0类患者的肝脏功能最佳，第2类患者明显不如第0类患者，与实验室检查指标得到的结果保持一致。

常规心脏超声指标包括心脏心房心室壁厚度，两心室收缩和舒张指标以及VFM相对压和能量损耗。通过血流向量成像技术可以获得患者的血液流动压强和能量损失，根据其医学意义，能量损失值越低，心脏供越好。通过比较三类聚类类别中指标，考虑体表面积标化的舒张快速充盈期的能量损耗，第0类的数据均值为14.50±16.07，第1类为16.47±20.21，第2类为19.10±19.58，则能够判断出三类类别的心脏功能等级依旧是按照0、1、2依次递减。

## 4.5 拓扑聚类算法和自编码器算法比较

对于自编码器深度聚类的K - S检验在上一章节中已经阐述，接下来进行拓扑聚类算法的检验。根据拓扑聚类模型得到的预测标签y_pred，分别进行三类类型两两间的检验，第0类和第1类样本的有效区分维度为：中性粒细胞、肝脏联合弹性指标LFI以及VFM-相对压收缩射血期IVPD；第0类和第2类样本的有效区分维度为：VFM-相对压等容收缩期IVPG；第1类和第2类样本的有效区分维度为：超声肝病分类、中性粒细胞、肝脏联合弹性指标FIndex（Med）。通过和自编码器深度聚类算法的对比，可以发现样本间的有效区分维度数量相

对较少，并且没有人何以各维度的信息能够对三个类别进行有效区分，因此判断自编码器深度聚类算法的聚类性能优于拓扑聚类算法。

经过对实验结果的分析，发现在进行拓扑聚类算法的时候，有一些数据因为高维开方体在对数据矩阵进行覆盖的时候，既被上一覆盖选中，又被下一覆盖选中，如下图中红圈部分所保全的数据样本点。这样在经过算法的训练之后，会发现有一些样本数据同时出现在了算法生成的拓扑图中两个拓扑点里。因此，再通过拓扑图的结构划归拓扑点的时候，一些数据就会被分到两个不同的类别中。对于这些被重复预测的数据，算法给出的做法是将这些数据样本点随机分配到两个类别中，从而影响算法对患者样本所属标签的判断，因此产生了差异。

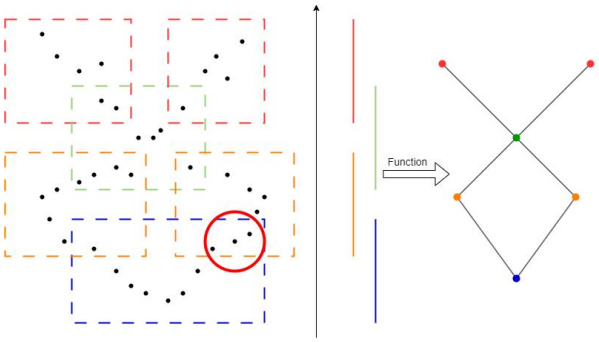


图 4-9 重复预测数据样本点

## 4.6 本章小结

本章节实现了自编码器聚类算法，具体实现了两种数据升维方式，三种自编码器神经网络以及讨论了初始化聚类中心个数数量。经过本章节的对比实验，最终确定聚类效果最好的自编码器网络架构为线性插值算法升维，Conv2DTranspose 方法上采样，聚类中心初始化个数为 3。经过自编码器的训练和深度聚类层的聚类，最终将 208 个数据样本分为三类，并对聚类结果进行了医学合理性分析：从实验室指标、肝功能对心功能的影响以及常规心脏超声指标三个方面验证了聚类结果的合理性，说明了自编码器深度聚类算法能够将患者样本进行合理的聚类并且得到符合医学解释的结果。

# 结　论

本文使用自编码器深度聚类算法对患者样本进行了心脏功能评级，实现了不同神经网络结构的自编码器算法，并使用哈尔滨医科大学提供的脱敏实验数据对神经网络模型进行训练，得到了相应的聚类结果，对患者的心功能进行了评级，并进行了医学合理性分析。

本文的主要工作和得出的结论如下：

1）基于拓扑聚类算法，使用脱敏后的患者心脏功能相关体检指标对患者样本进行了聚类，得到 3 类患者心功能划分，并结合了 K - S 检验以及医学合理性分析了每一聚类类别的心功能情况，总结出该模型聚类结果具有一定合理性；

2）实现了自编码器深度聚类算法，并对最优的自编码器结构进行了较为深入的探讨，最终得到最优自编码器模型为 Conv2DTranspose 卷积自编码器。基于该自编码器结构，通过训练得到 3 类患者心功能聚类类别，并对聚类结果进行了 K - S 检验和医学分析，验证了自编码器模型聚类结果的有效性和合理性；

3）对拓扑聚类算法和自编码器深度聚类算法得到的结果进行对比，分析了两种模型的优劣以及原因，最终确定基于自编码器深度聚类算法得到的聚类结果优于拓扑聚类算法，并总结出效果最好的聚类模型为 Conv2DTranspose 卷积自编码器。

本文的工作也有一定的不足之处，需要在日后的工作中进行深入的研究：在进行聚类中心初始化的过程时，由于使用 KMeans 算法，聚类中心值的选择会对最后得到的聚类结果产生较大的影响，有一定的不稳定性；样本数量较少，对于神经网络中的参数学习可能还没有达到极限。在未来的工作中，通过提升样本的数量以及优化初始化方法能够使模型更加稳定，最终得到更加合理的预测结果。本文的工作在通过扩充数据量以提升了模型稳定性后，可以更加广泛的应用于临床医学分析，通过将患者的心功能相关体检数据直接放入模型中来判断患者的心功能属于哪一个聚类分析，以达到提供医学辅助建议的作用。

# 参考文献

[1] Hamza El Hadi, Angelo Di Vincenzo, Roberto Vettor, Marco Rossato. Relationship between Heart Disease and Liver Disease: A Two-Way Street[J]. Cells. 2020：5-10.

[2] 樊弘,左丹,蒋芳萍. 慢性心力衰竭左室超声测量参数与心功能分级的关系研究. [J]川北医学院学报. Volume 37, Issue 12. 2022：1547-1549.

[3] Møller Søren, Bernardi Mauro. Interactions of the heart and the liver. [J]Journal European heart journal. Volume 34, Issue 36. 2013：2804-11.

[4] Zhang Yaxing, Fang Xian Ming. Hepatocardiac or Cardiohepatic Interaction: From Traditional Chinese Medicine to Western Medicine.[J]Evidence-Based Complementary and Alternative Medicine. Volume 2021, Issue 2021. 2021：1-14.

[5] 张辰宇, 穆心苇. 人血白蛋白在心脏加速康复外科中的作用. [J]Chin ECC Volume 21, Issue 3. 2023：179-182.

[6] Farzaneh Tajdini, Mohammad-Javad kheiri. Recent advancement in Disease Diagnostic using machine learning: Systematic survey of decades, comparisons, and challenges[cs.CV]. Volume 2308, Issue 1319. 2023：2-6.

[7] Jeffrey S Bennett, David M Gordon, Uddalak Majumdar, Patrick J Lawrence, Adrianna Matos Nieves, Katherine Myers, MS, Anna N Kamp, Julie C Leonard, MD,Kim L McBride, Peter White, and Vidu Garg. Use of Machine Learning to Classify High Risk Variants of Uncertain Significance in Lamin A/C Cardiac Disease. [J]Heart Rhythm. Volume 19, Issue 4. 2022：676–685.

[8] Jabir Al Nahian1, Abu Kaisar Mohammad Masum1, Sheikh Abujar2, Md. Jueal Mia. Common human diseases prediction using machine learning based on survey data. [J]Bulletin of Electrical Engineering and Informatics. Volume 11, Issue 6. 2022：1111-1121.

[9] Isra'a Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azzeh. A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods. [IJCSIS]International Journal of Computer Science and Information Security . Volume 14, Issue 12. 2016：868-879

[10] Ibrahem Alarsan, Fajr Ibrahem Alarsan. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. [J]Alarsan and Younes J Big Data. Volume 6, Issue 81. 2019：1-15.

[11] MacQueen J. Some methods for classification and analysis of multivariate observations. [C]Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Volume 1, Issue 14. 1967：281-297.

[12] Chazal F, Michel B. An introduction to topological data analysis: fundamental and practical aspects for data scientists. [J]Frontiers in artificial intelligence. Volume 4, Issue 108. 2021：1-44.

[13] Michelucci,Umberto. An Introduction to Autoencoders. [cs.LG]ArXiv. Volume 2201, Issue 3898. 2022：1-26.

[14] Chazal, Frédéric and Bertrand Michel. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. [J]Frontiers in Artificial Intelligence. Volume 4, Issue 4. 2021：1-44.

[15] Breiman, L. Random Forests. [J]Machine Learning. Volume 45, Issue 45. 2001：5-32.

[16] Jian Yang, D. Zhang, A. F. Frangi and Jing-yu Yang. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. [J]IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 26, Issue 1. 2004：131-137.

[17] Rosenblatt, Frank. The perceptron: a probabilistic model for information storage and organization in the brain. [J]Psychological review.Volume 65, Issue 6. 1958：386-408.

[18] Wang, Zhiguang and Tim Oates. Imaging Time-Series to Improve Classification and Imputation. [J]International Joint Conference on Artificial Intelligence. 2015：1-7.

[19] Ronen, Meitar et al. DeepDPM: Deep Clustering With an Unknown Number of Clusters. [CVPR]IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022：9851-9860.

[20] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. [J]In

Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996：226–231.

[21] Ruder, Sebastian. An overview of gradient descent optimization algorithms. [cs.LG]ArXiv. 2016：1-14.

[22] Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. [J]Information Sciences. 2023：178-210.

[23] 孙海娇,邱仕君. 试从阴阳、气血、经络论《黄帝内经》中心肝关系.[J]辽宁中医药大学学报. 2012：110-112.

# 哈尔滨工业大学本科毕业论文（设计）
# 原创性声明和使用权限

## 本科毕业论文（设计）原创性声明

本人郑重声明：此处所提交的本科毕业论文（设计）《基于自编码器聚类算法的心脏功能评级算法研究》，是本人在导师指导下，在哈尔滨工业大学攻读学士学位期间独立进行研究工作所取得的成果，且毕业论文（设计）中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本毕业论文（设计）的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：　　日期：　　年　　月　　日

## 本科毕业论文（设计）使用权限

本科毕业论文（设计）是本科生在哈尔滨工业大学攻读学士学位期间完成的成果，知识产权归属哈尔滨工业大学。本科毕业论文（设计）的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存本科生上交的毕业论文（设计），并向有关部门报送本科毕业论文（设计）；（2）根据需要，学校可以将本科毕业论文（设计）部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）本科生毕业后发表与此毕业论文（设计）研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉本科毕业论文（设计）的使用权限，并将遵守有关规定。

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

# 致　谢

毕设课题工作至此结束，也昭示了我的本科大学生活即将落幕。时间如白驹过隙般，转眼间我已经在哈尔滨工业大学度过了四年的时光，岁月如梭，慨叹万千。这段时光少不了家人、老师和同学们的帮助，在此，我想感谢在我本科生活中帮助过我的所有有人。

首先感谢我的毕设指导老师袁永峰老师和同组老师李钦策老师的帮助。在我进行毕设课题相关实验遇到自己解决不了的问题时，老师们都会给予我及时的帮助，对我的问题也会及时的反馈和解答，给出相关的建议，使我受益良多。我的毕业设计能够按时完成并得到现在的结果离不开老师的指导和帮助。

其次要感谢我的父母。上大学期间难免会遇到挫折和困顿，在我迷茫踌躇不行之时，父母的劝说和引导无疑给我了鼓励和精神支持，让我在遭受挫折之时能够快速回复，重回生活正轨。

最后我还想感谢我的母校哈尔滨工业大学。感谢母校为我提供了一个学习氛围好、成长机会多的环境。无论是学术还是个人方面，我都感到非常幸运能够在这样一个温馨而充满活力的校园中学习。再次感谢母校提供的付出和关怀，我将永远珍惜这段美好的时光。