2023-2024 学年秋季学期

《数据科学与工程 1-数据挖掘》3XS1080002

主观题答卷

学　　号	23724590	学　　院	材料基因组工程研究院	
姓　　名	郎志远	手工签名		
成　　绩	选题(10)	格式(10)	内容(40)	成绩
教师评语 教师签名 批阅日期　2023 年　11　月　　日				

基于多层感知机的出血性脑卒中患者预后预测

郎志远[1]

[1]（上海大学材料基因组工程研究院上海）

Prognostic prediction of hemorrhagic stroke patients based on multilayer perceptual machine

LangZhiyuan[1]

[1]（InstituteofMaterialsGenomeEngineering,ShanghaiUniversityShanghai）

1 引言

出血性脑卒中是指脑实质内血管非外伤性破裂导致的脑出血，占所有脑卒中病例的 10-15%。这种疾病的病因非常复杂，通常由于脑动脉瘤破裂、脑动脉异常等原因导致血液涌入脑组织，使脑部受到机械性损伤并引发一系列生理病理反应。出血性脑卒中脑具有突然发病、病情加重迅速、预后不佳等临床特征，急性期病死率高达 45-50%，并且患者会遭受严重神经功能障碍等后遗症。出血性脑卒中不仅让患者承受死亡的风险，还给社会与患者家庭带来沉重的健康和经济负担。因此，深入研究出血性脑卒中的发病风险，整合影像学数据换着临床信息及临床诊疗方案，使患者的预后预测更加精准，并以此为依据针对临床决策进行优化，具有重要的临床价值。在出血性脑卒中后，血肿范围扩大是预后不良的重要危险因素之一。短时间内，由于脑组织受损和炎症反应等因素，血肿范围可能逐渐扩大，导致颅内压急剧升高，进一步损害神经功能，甚至威胁患者生命。因此，需要重点检测和控制血肿的扩张情况。另外，血肿周围的水肿在近年来引起广泛关注，作为脑出血后继发性损害的标志，血肿周围水肿可能导致脑组织受到压迫，进而影响神经元功能，加重脑损伤，更进一步损害患者的神经功能。所以对血肿扩展和血肿周围水肿的早期识别和预测有着非常重要的价值。医学影像技术能够帮助我们动态监测出血性脑卒中后脑组织损伤，同时我们可以使用人工智能技术对大量影像数据进行数据挖掘与智能分析。借助脑出血患者的影像信息，结合患者的个人信息、治疗方案和预后数据，我们可以构建智能诊疗模型，明确导致出血性脑卒中预后不良的危险因素，实现精确的个性化治疗效果评估和预后预测。

2 问题定义

2.1 血肿扩张风险相关因素探索

根据患者的临床信息、患者血肿体积及位置影像信息判断患者发病后 48 小时内是否发生血肿扩张事件。并以是否发生血肿扩张事件为目标变量基于患者的个人史，疾病史，治疗方案以及患者血肿的首次影像检查结果构建模型，预测所有患者发生血肿扩张的概率构建模型，预测所有患者发生血肿扩张的概率。

2.2 血肿周围水肿体积情况分析及患者预后预测

根据患者的水肿体积和重复检查时间点，构建不同类别患者水肿体积随发病至影像检查时间进展曲线。利用患者的个人史，疾病史，治疗方案和首次影像检查结果等数据构建预测模型，预测患者的 90 天 mRS 评分。

3 数据集简介

本文使用的数据集涵盖 160 例出血性脑卒中患者的相关信息，包括三个表。患者列表及临床信息表中包含患者的个人史、疾病史、发病及治疗相关信息、多次重复的影像学检查结果及患者预后评估；患者影像信息血肿及水肿的体积及位置表中包含了患者的每个时间点血肿总体积及水肿总体积及不同位置的占比；患者影像信息血肿及水肿的形状及灰度分布表中包含每个时间点血肿及水肿的形状及灰度特征，反映目标区域内体素信号强度的分布（17 个字段）及三维形状的描述（14 个字段）

4 数据挖掘方法

4.1 数据预处理

4.1.1 数据清洗

在患者在重复进行影像检查不能保证每次影像检查的数据均能反映水肿、血肿体积等相关特征的实际情况，因此数据中可能同时存在真实数据和具有干扰信息的异常数据，针对具有干扰信息的异常数据，我选择基于 Z-Score 对异常数据进行清洗。Z-Score 是一种常用于检测和清洗异常数据的统计方法。Z-Score 的计算可以帮助我们理解数据点相对于整个数据集的位置，以及它们的离散程度，它通过测量数据点与其均值的偏差来标准化数据，并将其表示为标准差的倍数。我们首先计算每个数据点的 Z-Score，将 Z-Score 的绝对值大于某个阈值的数据点定义为异常值，根据阈值来清洗异常数据，可以选择将异常值从数据集中删除。数据清洗前后患者水肿体积随时间变化的离散图如图 1 所示

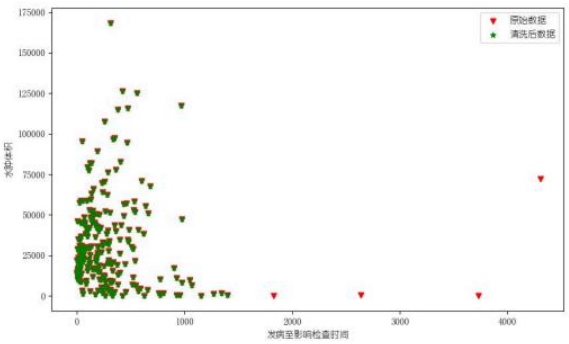

图 1 数据清洗前后患者水肿体积随时间变化对照

4.1.2 特征重编码

由于数据集中不同的治疗方案对应了数据集的多个特征，考虑到治疗方案之间的组合方式我们可以将治疗方案进行重编码，将其由多个特征组合为一个特征。本文采用二进制编码的方式将多个治疗方案特征重新进行编码，使用二进制编码中，每个方案特征都会被映射到一个二进制编码，其中每个二进制位代用 1 和 0 分别代表采用与不采用该种治疗方案，使用一串二进制编码可以有效地表示多个治疗方案的组合，并且不引入治疗方案之间的顺序关系。

4.2 曲线拟合

4.2.1 多项式曲线拟合

多项式曲线拟合是一种用多项式函数来逼近或拟合一组离散数据点的方法。通过选择合适的多项式阶数和系数，使得多项式曲线能够最好地拟合给定的数据点。多项式函数通常表示为如下形式：$f(x)=a_0+a_1x+a_2x^2+\cdots+a_nx^n$其中，$f(x)$是多项式函数，x 是自变量，$a_0,a_1,a_2,\ldots,a_n$是多项式的系数，n 是多项式的阶数。每个系数对应于多项式中的一个幂次。考虑有 m 个数据点(x_i,y_i),需要找到多项式的系数$(a_0,a_1,a_2,\ldots,a_n)$使得多项式曲线尽可能地接近这些数据点。为每一个数据点构建一个线性方程：$y_i=a_0+a_1x_i+a_2x_i^2+\cdots+a_nx_i^n$这样就得到了一个包含 m 个线性方程的系统。将这些线性方程组合成矩阵形式：Y=XA。在此式中 Y 为包含所有y_i的列向量，X 为 m×(n+1)的矩阵，每一行对应于一个数据点，每一列包含对应x_i的不同幂次的项，A 是包含多项式系数的列向量。为了找到最佳的多项式系数 A，可以使用最小二乘法来求解。

4.2.2 高斯曲线拟合

高斯曲线拟合是一种常用于数据分析和模型建立的方法，它基于正态分布（也称为高斯分布）的数学性质，通过找到一条或多条高斯曲线，最佳地拟合给定的数据分布。对一组数据(x,y)(i=1,2,3…)使用高斯函数进行描述：

$$y_i=y_{max}\times e^{\frac{(x_i-x_{max})^2}{S}} \qquad (1)$$

(1) 式两边同时取对数可化为：

$$\ln y_i=\ln y_{max}-\frac{(x_i-x_{max}^2)^2}{S}=\left(\ln y_{max}-\frac{x_{max}^2}{S}\right)+\frac{2x_ix_{max}}{S}-\frac{x_i^2}{S} \qquad (2)$$

令

$$\ln y=z,\ \ln y_{max}-\frac{x_{max}^2}{S}=b,\frac{2x_{max}}{S}=b_1,-\frac{1}{S}=b_2 \qquad (3)$$

对于全部实验数据，以矩阵的形式可以将(2)式表示为：

$$\begin{bmatrix}z_1\\z_2\\z_3\\\cdots\\z_n\end{bmatrix}=\begin{vmatrix}1&x_1&x_1^2\\1&x_2&x_2^2\\1&x_3&x_3^2\\\cdots&\cdots&\cdots\\1&x_n&x_n^2\end{vmatrix}\begin{bmatrix}b_0\\b_1\\b_2\end{bmatrix} \qquad (4)$$

记为：Z=XB，成矩阵 B 的广义最小二乘解为：

$$B=(X^TX)^{-1}X^TZ$$

根据（3）式求出待估参数，得到高斯函数的特征参数

4.3 聚类划分亚组

4.3.1 高斯混合聚类

高斯混合模型（GaussianMixtureModel，GMM）是一种基于概率分布的聚类方法，通过用多个高斯分布函数去近似任意形状的概率分布，假设数据是由多个高斯分布组成的混合体。每个高斯分布成为一个“分量”，然后找到这些高斯分布的参数以及每个数据点属于每个分量的概率。GMM 假设数据是由多个高斯分布组成的混合体，分量数表示高斯分布的个数。这个数量通常需要根据数据的性质来选择，可以是事先确定的，也可以通过模型选择方法估计得出，每个高斯分布都有其均值和协方差矩阵作为参数，用于描

述分布的形状和位置使用混合系数代表每个高斯分布的权重，表示每个分量在整个数据集中的相对重要性，这些权重之和等于 1。GMM 的聚类过程通常包括以下步骤：

1.初始化：随机初始化每个高斯分布的均值、协方差矩阵和混合系数。

2.使用 Expectation-Maximization(EM)算法：来迭代地估计模型参数。EM 算法包括两个步骤：a.Expectation 步骤：使用贝叶斯定理和高斯分布的概率密度函数来计算每个数据点属于每个分量的后验概率，即计算每个数据点在每个高斯分布下的概率权重。b.Maximization 步骤：在这一步中，根据 Expectation 步骤得到的后验概率，更新每个高斯分布的均值、协方差矩阵和混合系数，以最大化似然函数。

3.收敛判定：通常设置参数变化的阈值或最大迭代次数为收敛条件，根据收敛条件检查模型参数是否已经收敛。

4.结果输出：一旦模型收敛，输出每个高斯分布的均值、协方差矩阵和混合系数，以及每个数据点属于哪个分量的后验概率，根据后验概率将数据点分配到对应的簇。

4.3.2K-Means 聚类

K-Means 聚类是一种常用的无监督学习算法，用于将数据集中的观测点划分为 K 个不同的簇，每个簇代表一个数据群体。该算法的目标是最小化数据点与其所属簇内的聚类中心之间的距离，从而实现簇内数据的相似性最大化，簇间数据的相似性最小化。K-Means 聚类是一种迭代算法，通过不断地分配数据点到最近的聚类中心并更新聚类中心，最终实现数据集的聚类分析。对于有一个包含 N 个数据点的数据集，其中每个数据点表示为 $x_1,x_2,...,x^N$，每个数据点属于 K 个簇之一，使用 K-Means 方法进行聚类的步骤如下：

1.初始化：选择 K 个聚类中心，通常表示为 $\mu_1,\mu_2,...,\mu_k$。这些聚类中心是从数据集中随机选择的，或者可以通过其他初始化方法得到。

2.分配数据点：对于每个数据点 x_i，计算其与 K 个聚类中心的距离，通常使用欧氏距离或其他距离度量方式。将数据点 x_i 分配到距离最近的聚类中心

4.4 多层感知机

多层感知机(MLP)是一种适用于多特征、多类别分类任务的神经网络结构。它由多个全连接层组成，具备处理复杂非线性关系的能力。本文采用的 MLP 模型包含三个全连接层。全连接层是神经网络的核心组件，每个神经元与前一层的所有神经元相连接，从而能够学习输入特征之间的复杂关系。在多层感知机的每个全连接层之间，使用 ReLU 激活函数，它引入了非线性性质，有助于模型学习非线性函数，在最后一个全连接层之后，使用 softmax 来计算概率，由于是二分类任务，输出层有两个神经元。对全连接层的权重进行 Xavier 初始化。Xavier 初始化有助于避免梯度消失问题，破坏权重的对称性，以及提高模型的训练速度和泛化能力。模型的损失函数选择了交叉熵损失函数，这是多类别分类任务中常用的损失函数。我们选择了 Adam 优化器，它是一种自适应学习率算法，能够自动调整学习率以最小化损失函数。模型通过迭代训练数据集来学习特征之间的关系和类别之间的决策边界。在每个迭代周期（epoch）中，计算损失并使用反向传播算法来更新权重，以不断优化模型，提高分类性能。为了提高模型的泛化能力和鲁棒性，我们引入了随机噪声进行训练。随机噪声可以提供正则化效应，帮助避免过拟合，并增加模型的探索性，以更好地泛化到未见过的数据。

5 数据挖掘过程及结果

首先根据患者相关信息数据集中的数据判断病人是否发生血肿扩张。利用发病到首次影像检查时间间隔和影像检查时间点，计算相对时间。相对时间表示了从发病到首次影像检查时间间隔加上随后的影像检查时间点距离入院首次检查 的时间差。这个相对时间将在后续步骤中用于判断每次影像检查是否在发病后 48 小时内，将对每一次影像检查数据进行分析，计算体积变化，并根据上述条件来判断是否发生了血肿扩张事件。得到患者中发生血肿扩张与未发生血肿扩张的比例如图 2 所示

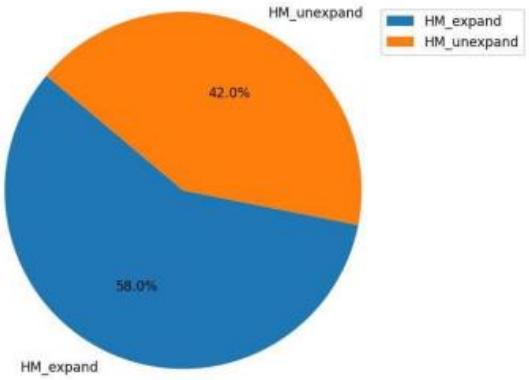

图 2 患者中是否发生血肿扩张人数比例

接着对血肿扩张事件进行预测，基于患者的个人史、疾病史、发病相关信息、治疗相关 特征以及首次影像检查结果等变量，构建一个二分类模型，用于预测患者是否会在未来发生血肿扩张事件。对数据进行异常值处理和归一化处理后使用多层感知机进行预

测分析。训练完成的模型的损失情况如图 3 所示

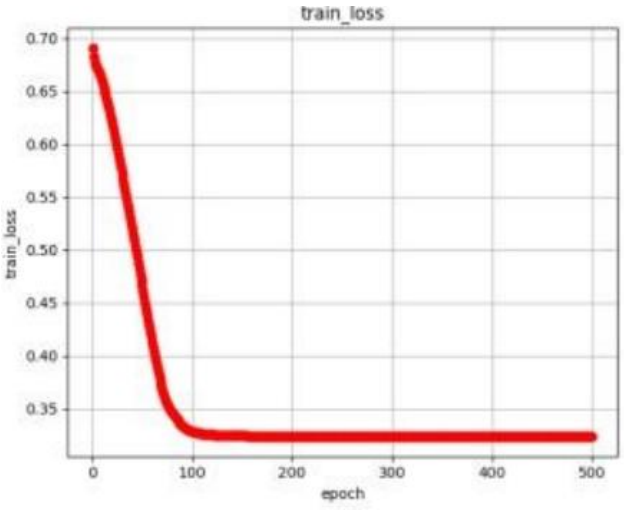

图 3 模型损失情况

训练完成的模型的准确率如图 4 所示：

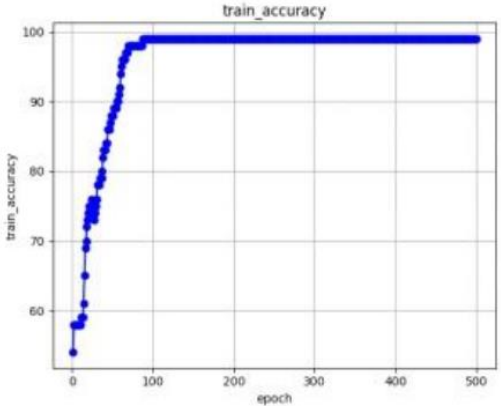

图 4 模型准确率

为探索患者水肿体积随时间进展模式存在的个体差异，首先将患者聚类分为不同亚组。使用 GMM 模型

将患者聚类为 4 个亚组的结果如图 5 所示：

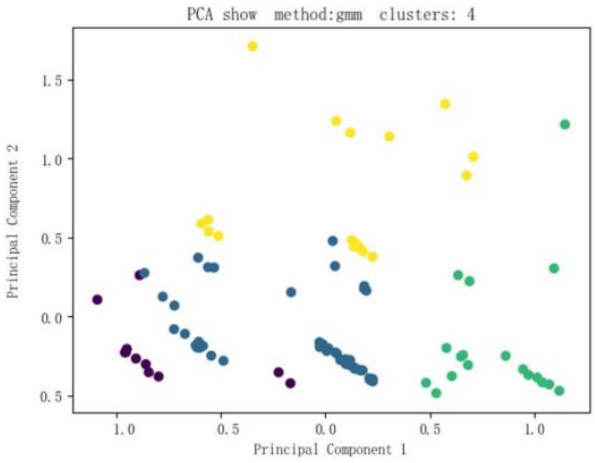

图 5 使用 GMM 进行聚类的结果

使用 K-Means 聚类模型将患者聚类为 4 个亚组的具体效果如图 6 所示：

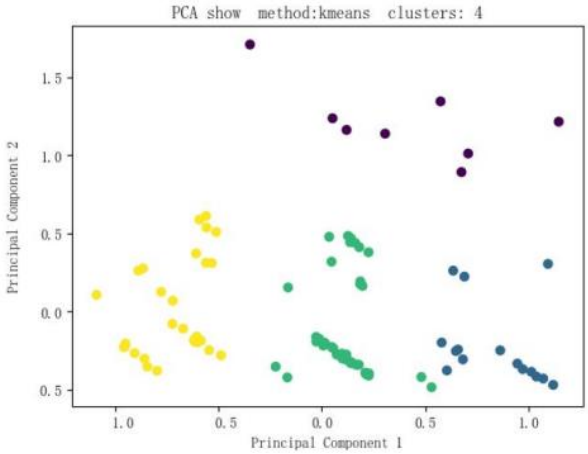

图 6 使用 K-Means 进行聚类的结果

对比之下使用 K-Means 聚类模型时所划分的亚组在空间中更为集中，联系更加紧密，聚类效果要比 GMM 模型更好。

聚类完成后不同人群的水肿体积随时间进展曲线进行拟合的结果如图 7 所示：

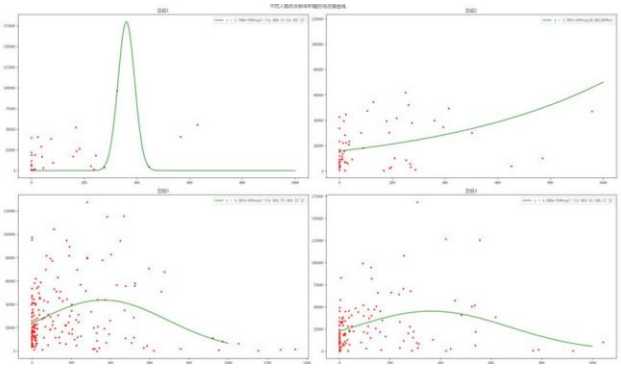

图 7 不同人群的水肿体积随时间进展曲线进行拟合的结果

最后根据前 100 位患者的个人史、疾病史、发病相关信息以及首次影像结果，构建一个预测模型，用于预测所有患者在 90 天后的 mRS 评分。首先我首次影像检查结果中的特征，构建一个回归任务，预测取值范围在 0 到 6 之间的有序等级变量，代表不同的 mRS 评分等级。在预测时仍然使用多层感知机作为模型架构，训练过程中随 epoch 变化模型的准确率变化如图 7 所示：

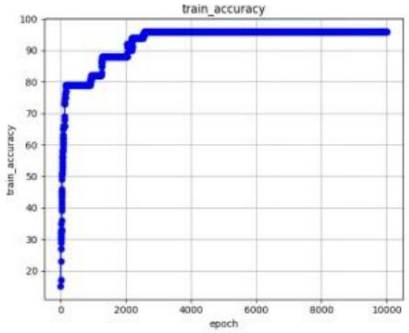

图 8 随 epoch 变化对应的模型准确率

保存准确率最高模型训练参数后，使用该参数对目标值 mRS 进行预测。

6.分析与讨论

本文在进行数据挖掘的过程中充分结合实际，简化脑卒中预测条件，考虑了诸多重要因素得到合理的模型，如患者的年龄、性别、疾病史等。这样得到的模型贴合实际，具有较高的应用价值，可以推广到不同患者群体。构建的模型运用深度学习和神经网络思想，抓住影响脑卒中预测问题的重要因素，将复杂的医学数据转化为简单的预测问题，合理设置参数，模型的输出结果符合医学需求，能解决实际的预测问题。本文使用的多层感知器算法具有非线性建模能力、自适应学习率等优点，对于求解脑卒中预测模型非常适用。但是在实际应用中，患者的遗传因素和生活方式可能也是重要的因素，但本文未能考虑到这些因素的影响，一定程度上影响了模型的准确性；实际上，患者的年龄、疾病史等因素的影响可能不一定是线性的，而本文将其作 为线性因子处理，忽略了边际效应的影响。在模型推广方面，可以将模型参数进行调整，例如，结合遗传信息等因素，从而 解决遗传性脑卒中患者的预测问题。改进方面，可以结合参考文献中的新研究，进一步考虑患者的生活方式、遗传信息等对脑卒中预测的影响，从而得到更全面、更合理的模型。这些改进将有助于模型在更广泛范围内的应用，提高脑卒中患者的预测准确性，为临床决策提供更多有力的支持。

7.收获和体会

在本次数据挖掘过程中，我深刻领悟到了数据挖掘的重要性和应用价值。以下是我在这个过程中的一些心得体会：

1. 数据准备是关键：在进行数据挖掘之前，充分准备和清洗数据是至关重要的步骤。数据的质量和完整性直接影响到挖掘结果的准确性和可靠性。因此，我学会了如何对数据进行预处理、去除噪声、处理缺失值和异常值等操作，以确保数据的准确性和可靠性。

2. 特征选择和特征工程的重要性：在进行数据挖掘任务时，选择合适的特征对结果具有重要影响。通过对数据进行特征选择和特征工程，可以提取出最具判别性和相关性的特征，从而提高模型的性能和准确性。

3. 模型选择和评估：在进行数据挖掘时，选择适合问题的模型非常重要。不同的问题可能需要不同类型的模型，如分类、聚类、回归等。我学会了评估不同模型的性能，并选择最适合问题的模型。

4. 结果解释和可视化：数据挖掘的结果往往需要解释和呈现给他人。我学会了如何解释模型的结果并将其可视化，通过可视化，能够更清晰地理解数据的模式和关联，并向他人有效地传递这些信息。

5. 持续学习和改进：数据挖掘是一个不断发展和演进的领域，新的算法和技术不断涌现。在进行数据挖掘的过程中，我意识到持续学习和改进是非常重要的。

参考文献

[1] 朱遂强,刘鸣,等.中国脑出血诊治指南(2019)[J].中华神经科杂志,2019,52(12):12.DOI:10.3760/cma.j.issn.1006-7876.2019.12.003

[2] 张丽娜,李国春,周学平等.基于支持向量机的急性出血性脑卒中早期预后模型的建立 与评价[J].南京医科大学学报(自然科学版),2016,36(01):80-84.

[3] Hamerly G, Elkan C. Learning the k in k-means[J]. Advances in neural information processing systems, 2003, 16.

[4] Hamerly G, Elkan C. Learning the k in k-means[J]. Advances in neural information processing systems, 2003, 16

[5] Al Shalabi L, Shaaban Z, Kasasbeh B. Data mining: A preprocessing engine[J]. Journal of Computer Science, 2006, 2(9): 735-739

[6] Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11).

[7]] Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels[J]. Advances in neural information processing systems, 2018, 31.

[8] Noriega L. Multilayer perceptron tutorial[J]. School of Computing. Staffordshire University, 2005, 4(5): 444

[9] Bro R, Smilde A K. Principal component analysis[J]. Analytical methods, 2014, 6(9): 2812-2831

[10] Noriega L. Multilayer perceptron tutorial[J]. School of Computing. Staffordshire University, 2005,