模糊系统大作业

指导老师：郭大蕾

姓名：陈帅华

学号：202028014728006

目录

[模糊聚类过程中的方法介绍 3](#_Toc55223364)

[鸢尾花数据集的预处理 3](#_Toc55223365)

[关于聚类后对聚类中心类别的判断 3](#_Toc55223366)

[关于聚类时训练误差的计算 3](#_Toc55223367)

[关于聚类后准确率的计算 4](#_Toc55223368)

[模糊聚类的结果与讨论 4](#_Toc55223369)

[原始数据可视化 4](#_Toc55223370)

[聚类结果可视化 5](#_Toc55223371)

[训练误差与迭代次数间的关系 6](#_Toc55223372)

[不同聚类数聚类准确率的比较 7](#_Toc55223373)

[与BP神经网络结果的比较分析 8](#_Toc55223374)

# 模糊聚类过程中的方法介绍

## 鸢尾花数据集的预处理

鸢尾花数据集中共有三类鸢尾花，分别为：setosa, versicolor,virginica。每一类鸢尾花有50条数，共有150条数据，而每一条鸢尾花数据均有四个特征，分别为：sepal length, sepal width, petal length, petal width。

在本次作业中，我们用1，2和3分别作为这三类鸢尾花的标签，即1代表setosa，2代表versicolor，3代表virginica。在聚类过程中，我们同时使用了这四个特征。

对于鸢尾花数据集中的150条数据，我们首先将其随机打乱，然后以7:3的比例，将整个数据集分为训练集和测试集两类，即训练集中105条数据，测试集中45条数据；其次根据训练集中的数据来进行模糊K均值聚类(聚类个数为3，4或5)。

## 关于聚类后对聚类中心类别的判断

在使用模糊K均值聚类聚类后，我们可以得到一个n×m维的聚类中心矩阵center，其中n代表聚类的个数，而m代表在聚类时使用的特征数。由于在聚类后，我们并不能确定center的每一行分别代表着鸢尾花的哪一类，因而也就不能根据隶属度矩阵U来确定聚类后每一类的个数。故而本作业采取了如下的方法：

当聚类个数等于3时，仅使用上述的四个特征来进行聚类，在计算得到聚类中心之后，我们通过聚类中心center(3×4的矩阵)第一列(即setosa这一特征)中三个数值大小来进行判断每一个聚类中心代表哪一类，例如最大的即代表virginica，最小的即代表setosa，中间值即代表versicolor。之后便可根据隶属度矩阵U(3×105)来判断训练数据中的每一条数据分别归属于哪一类，即如果某一条数据对于其中一种鸢尾花的隶属度大于0.5，就将它归为这一类。

当聚类个数等于4或5时，便于聚类个数等于3时不同，我们同时将上述的四个特征加上数据的标签来进行聚类，最终得到的聚类中心矩阵center维数变为4×5或者5×5。这时我们更不能直接来判定其中一个聚类中心代表着哪一类的鸢尾花。在本作业中，对center中最后一列数据进行了向距离最近的整数进行了一个取整，然后根据取整得到的值来判定每一个聚类中心代表哪一类鸢尾花。接着采取同样的方法根据隶属度矩阵来判断训练数据中的每一条数据分别归属于哪一类。

## 关于聚类时训练误差的计算

在模糊K均值聚类函数内部，每进行一次迭代，都可以得到一个聚类中心矩阵center和隶属度矩阵U。每得到一个center和U，我们都采取在上一部分中叙述的方法，来对训练集中每一条数据所属类别的划分。之后我们便可以计算每一次迭代得到的均方误差(MSE,Mean Squared Error)，其计算公式如下：



其中，yreal代表每一条数据实际的类别标签，而ypredict代表每一条数据聚类之后预测的类别类别标签。

## 关于聚类后准确率的计算

通过在上面论述中提及的方法，我们就完成了对聚类后聚类中心(模糊K均值聚类函数最终输出的聚类中心)所属鸢尾花类别的判断，也完成了聚类后训练集中每一条数据所属类别的划分。这样，我们便可计算出**训练准确率**。

在得到聚类中心的位置后，我们可以计算出测试集中每条数据距离每一个聚类中心的欧几里得距离，测试集的数据距离哪一个聚类中心最近，就将其归未该聚类中心对应的那一类鸢尾花。这样，我们便可计算出**测试准确率**。这样计算的话，就不存在测试误差，因而在本作业中，未讨论测试误差随迭代次数的变化。

对于不同的聚类个数(3,4或5)，我们都将程序运行三次，然后将得到的训练准确率和测试准确率取一个均值，作为最终的分类准确率。

# 模糊聚类的结果与讨论

## 原始数据可视化

我们将原始数据中的四个特征两两组合，然后将其分别作为横轴和纵轴绘制出了三种鸢尾花进行对比的六幅散点图，如图1所示。通过图1，我们可以发现，这三类数据在这四种特征上均有较大差异。

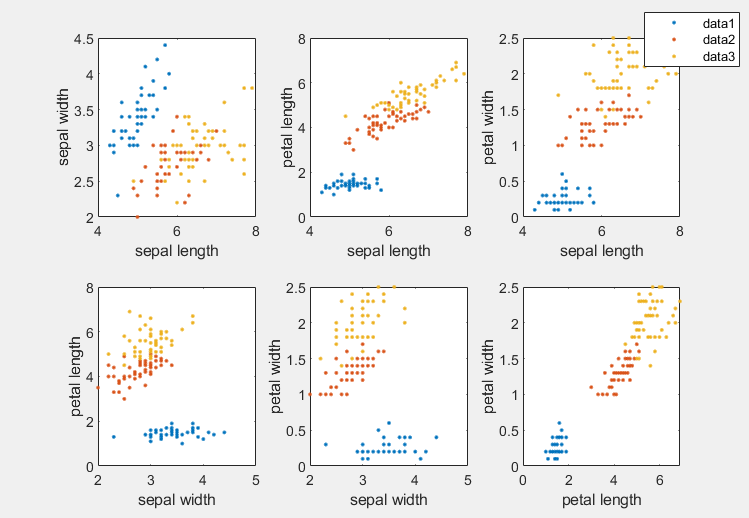


图1 原始数据散点图

## 聚类结果可视化

当聚类个数分别为3、4和5的时候，我们同样将四个特征进行两两组合，绘制出聚类结果的散点图，分别如图2、图3和图4所示。

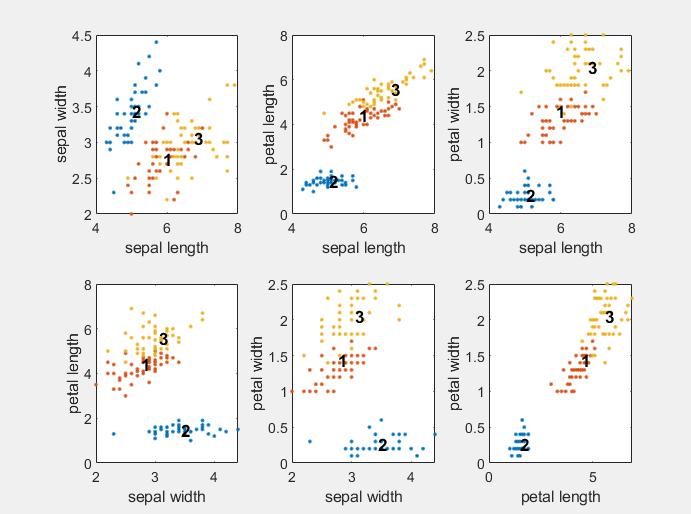


图2 聚类结果散点图(聚类个数为3)

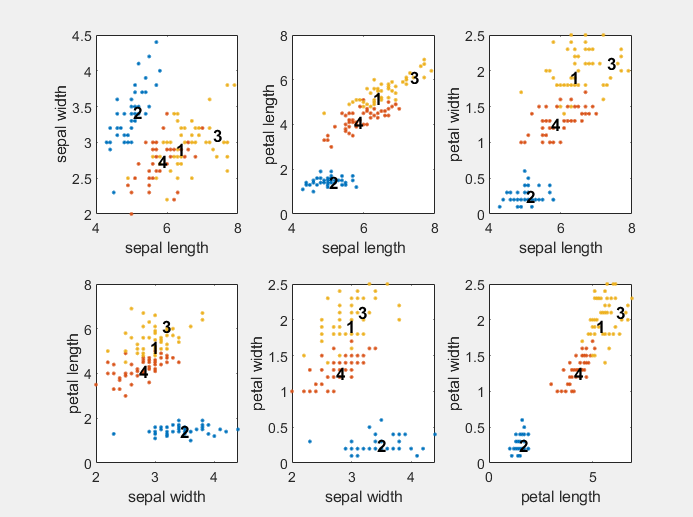


图3 聚类结果散点图(聚类个数为4)

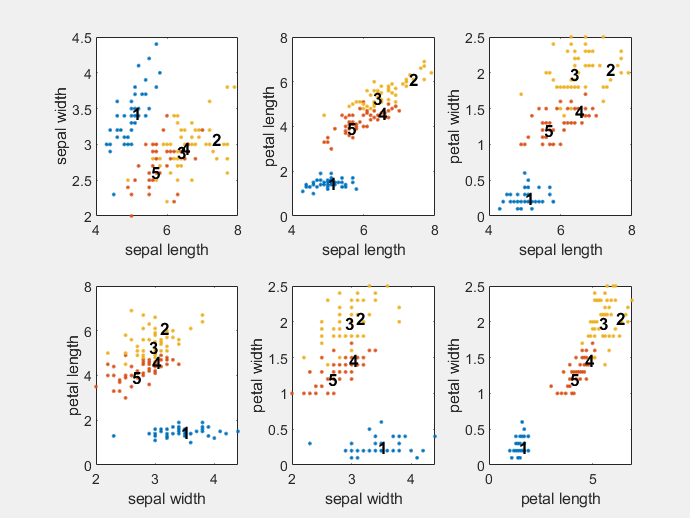


图4 聚类结果散点图(聚类个数为5)

## 训练误差与迭代次数间的关系

聚类个数为3、4和5的时候，训练误差随迭代次数变化曲线分别如图5、图6和图7所示。通过这三幅图像，可以发现，其最开始时的误差均在4点多，而且它们都是在很少的迭代次数中，均方误差就已经收敛到一个较小的值。

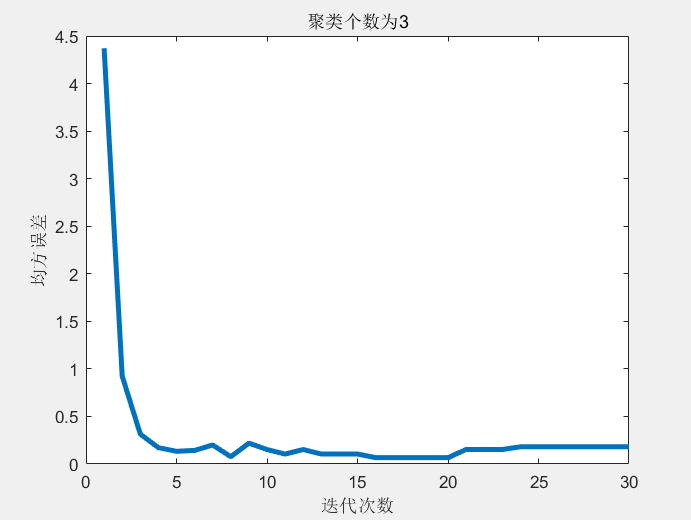


图5 训练误差随迭代次数变化曲线(聚类个数为3)

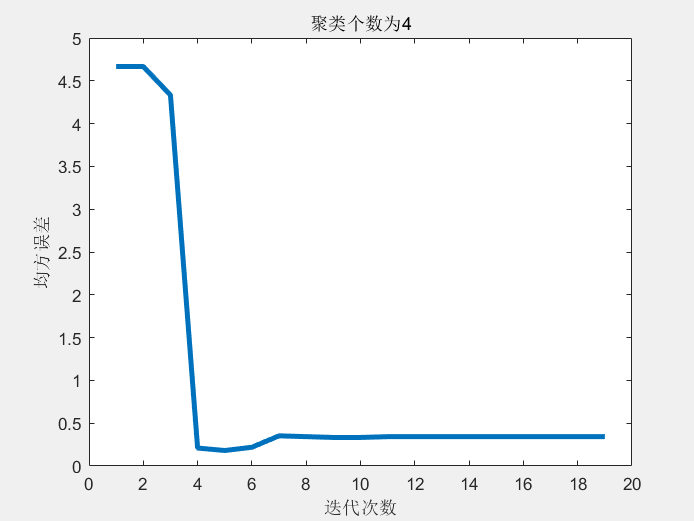


图6 训练误差随迭代次数变化曲线(聚类个数为4)

聚类个数为5时：

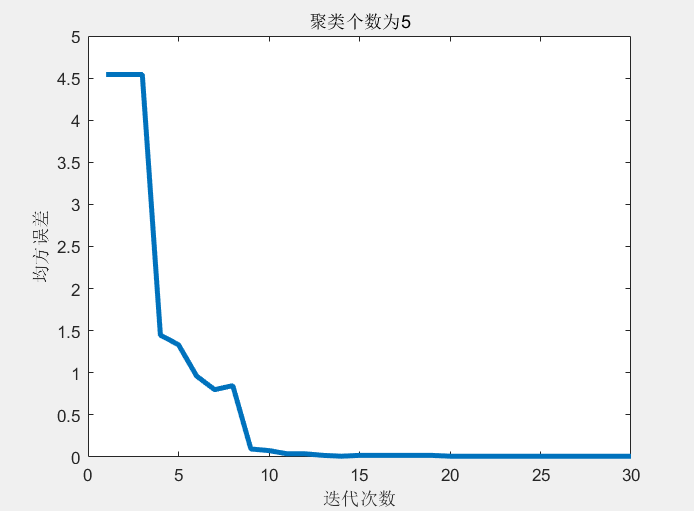


图7 训练误差随迭代次数变化曲线(聚类个数为5)

对于不同的聚类个数，在程序运行结束后得到的均方误差时大时小，无法找出误差随聚类数变化的规律，因此此处便不再对其进行分析。

## 不同聚类数聚类准确率的比较

图8中绘制了不同聚类数下训练准确率比较与测试准确率比较的柱状图。通过这幅图，我们可以发现，**不论是训练准确率还是测试准确率都随聚类数的增加而提高。**

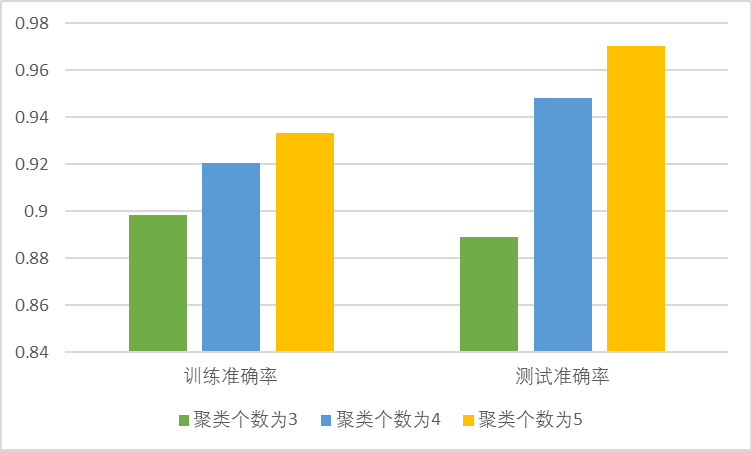


图8 不同聚类个数分类准确率的对比

## 与BP神经网络结果的比较分析

图9和图10分别是在BP神经网络中使用随机梯度下降法和动量法得到损失函数随迭代次数变化的曲线。将图5、图6和图7与图9和图10做一个对比(两者纵轴的含义不同，我们此处对比横轴)，我们可以发现使用模糊K均值聚类在很少的迭代次数(几十次)中就能使得误差收敛，而BP神经网络则需要上千次的迭代才能达到这一点。因而，可以说模糊K均值在这一点上优于BP神经网络。

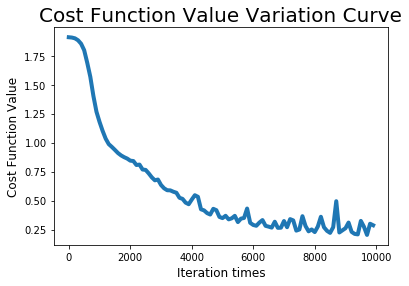


图9 SGD(hidden layer=10 learning rate=0.005)

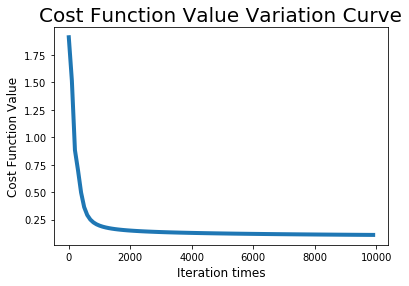


图10 动量法(hidden layer=10 learning rate=0.005)