数据挖掘过程报告作业三

王钰

2120171076

1. **问题描述**

从以下3个数据集中任选一个

* [https://www.kaggle.com/c/predict-west-nile-virus/data]
* [https://www.kaggle.com/c/predict-wordpress-likes/data]
* [https://www.kaggle.com/c/titanic/data]

**要求：**

1. 使用分类模型（至少2个）对数据集进行挖掘；
2. 对挖掘结果进行可视化，并解释其意义；
3. 使用聚类方法（至少2种）对数据集进行分析；
4. 对挖掘结果进行可视化，并解释其意义。

**提交的内容**

* 对数据集进行挖掘的源程序
* 挖掘结果及分析
* 挖掘过程的报告

1. 实验环境

系统：windows

编程语言：python

库：pandas、sklearn、graphviz（用来画决策树）

数据集：titanic[https://www.kaggle.com/c/titanic/data]

1. 实验分析
   1. 属性分析

PassengerID（ID）   
•Survived(存活与否)   
•Pclass（客舱等级，较为重要）   
•Name（姓名）   
•Sex（性别，较为重要）   
•Age（年龄，较为重要）   
•Parch（直系亲友）   
•SibSp（旁系）   
•Ticket（票编号）   
•Fare（票价）   
•Cabin（客舱编号）   
•Embarked（上船的港口编号）

通过对属性的分析，Sex ，Age, Parch，SibSp, Fare, Pclass会对是否存活有比较大的影响。所以，实验中选取了这6个属性及逆行挖掘。

* 1. 聚类

在聚类方法中，选取了（1）K-means算法：y\_pred = KMeans(n\_clusters=4, random\_state=0).fit\_predict(X\_train)（2）birch算法：y\_Birch = Birch(n\_clusters=None).fit\_predict(X\_train)

* 1. 分类模型

该实验选用了SVM和决策树（design tree）模型。