# 信用卡数据分析

吴杰德 2015311583

**摘要：**本文首先对信用卡数据进行数据预处理，将文本数据转换为数值型数据，并将数据进行z-score规范化。然后通过SVM-RFE特征筛选、相关性分析和Embedded特征筛选将筛选出最值得训练的26个数据特征。最后本文分别使用了6个分类训练模型和4个集成分类模型对信用卡数据进行训练和测试，结果表明单个分类模型中线性判别分析（LDA）和支持向量机（SVM）表现最好，准确度和F1得分都达到78%；集成分类模型中基于逻辑回归（Logistics Regression）、随机森林(Random Forest)和朴素贝叶斯（Naïve Bays）的投票分类器表现最好，F1 得分达到79%。

## 1. 数据预处理

### 1.1. 文本数据转换为数值数据

由于数据存在非数值型数据，故需要将文本型数据转换成数值型数据才能将该数据用于训练分类模型。因为非数值型数据存在不全为有序数据，故将文本型数据按无须数据处理，通过0,1哑变量，将文本数据转换为数值数据。

### 1.2数据规范化

数据每个特征的取值范围相差较大，直接用该数据对模型进行训练容易产生偏差，故将数据通过 Z-score 方法将数据规范化。

## 2. 特征工程

通过0,1哑变量将文本型数据转换为数值型数据后，数据特征存在冗余，因为需要对数据特征进行挖掘。

### 2.1. SVM-RFE特征筛选

SVM-REF的思想是根据SVM在训练时生成的权向量w来构造特征排序系数，每次迭代去掉一个排序系数最小的特征，最终得到所有特征属性的递减排序。

本文将变换后的61个特征通过SVM-RFE从中筛选30个特征。

### 2.2. 相关性分析

由于数据存在高度相关甚至完全正相关的特征，因此本文去除相关性高于70%的特征。通过相关性分析，本文去除了一个完全正相关的特征。

### 2.3. Embedded特征筛选

本文首先将数据进行随机森林训练，得到数据每个特征的重要程度，对于得分低于0.01的特征进行丢弃。本次筛选了3个特征。

最终，通过特征工程，筛选剩余的数据特征数为26个。

## 3. 模型训练

### 3.1. 数据划分

本文将数据划分成训练数据和测试数据，其中75%的数据用于训练模型，25%的数据用于测试模型。

### 3.2. 模型选择

对于信用卡状况分析的分别使用了单个分类模型和集成分类模型进行建模。

对于单个分类模型，本文使用了线性判别分析（LDA）、决策树（Decision Tree）、基于伯努利分布的朴素贝叶斯（Bernoulli Naïve Bayes）、最邻近（KNN）、逻辑回归（Logistics Regression）和支持向量机（SVM）。

对于集成模型，本文使用了极限随机森林（Extra Random Forest，ERF）、AdaBoost、梯度树提升（Gradient Tree Boosting）、基于逻辑回归、随机森林和朴素贝叶斯的投票分类器（Voting）。

### 3.3. 结果评价

表1. 模型测试结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 |
| LDA | 0.78 | 0.77 | 0.78 | 0.78 |
| Decision Tree | 0.668 | 0.71 | 0.67 | 0.68 |
| BernoulliNB | 0.776 | 0.8 | 0.78 | 0.78 |
| KNN | 0.716 | 0.71 | 0.72 | 0.71 |
| Logistic Regression | 0.772 | 0.77 | 0.77 | 0.77 |
| SVM | 0.784 | 0.77 | 0.78 | 0.78 |
|  |  |  |  |  |
| ERF | 0.772 | 0.77 | 0.77 | 0.77 |
| AdaBoost | 0.748 | 0.73 | 0.75 | 0.74 |
| Gradient Boosting | 0.76 | 0.75 | 0.76 | 0.76 |
| Voting | 0.784 | 0.79 | 0.78 | 0.79 |

表1展示了模型的测试结果。可以看出在单个分类模型中，线性判别分析、支持向量机表现最好，准确度和F1得分都达到78%，而在集成模型中，基于逻辑回归、随机森林和朴素贝叶斯的投票分类器表现最好，F1 得分达到79%。

F1 = 2\*precision\*recall / (precision + recall)

## 4. 结论

分类模型能较够较为准确的分析信用卡的状况，其中，对于单分类模型中，表现最好的是支持向量机（SVM），准确性和F1达到78%。另外通过投票的方式集成逻辑回归（Logistics Regression）、随机森林（Random Forest）和朴素贝叶斯（Naïve Bayes），能在一定程度上提高模型的性能，将F1提高到79%。