泰坦尼克事件获救人群分类问题

11510828 吕本猛

摘要

泰坦尼克沉船事件是上世纪震惊世界的一次灾难，引起了很大的轰动。在本问题中，要求根据每位乘客的信息变量，以及他们是否获救的情况，来对是否获救的人进行分类，以期望预测符合哪些条件的人会获救。

在建立模型进行训练和测试之前，先对所给的数据集进行数据预处理。这里将变量‘Sex’和‘Embarked’的数据改变为数字型编码，使其数据的形式变为数值形式。将变量‘Age’、‘Fare’利用平均数进行缺失值填补，将变量‘Embarked’利用众数进行缺失值填补。由于变量‘Ticket’的数据格式多种多样，不便于处理，变量‘Cabin’的缺失值过多，变量‘Sibsp’与变量‘Survived’之间的相关性很小，将这三列数据直接删除。之后，再对训练模型中要用到的变量与变量‘Survived’之间做相关性分析，得到在泰坦尼克事件中女性乘客有更大的几率获救，选择三等舱位的乘客有更大的几率不能获救，携带家属的乘客有更大的几率获救，支付的票价较高的人有更大的几率可以获救等规律。

在训练模型中，我们将数据分为训练集和测试集，其中测试集占30%。同时我们建立了多个不同的模型，分别使用 KNN 算法、支持向量机、朴素贝叶斯、决策树和随机森林算法来对数据进行训练，以求从多种不同的模型中获得一个准确度最高，效果最好的模型。最终根据不同模型的结果，随机森林模型可以更准确地预测出哪些人群会被获救，可以达到 90% 的准确率，最终能够在解决这一分类问题上达到很好的效果。

问题求解

数据预处理

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

图 1: 数据集

在导入数据之后，通过观察所给数据集的情况，我们发现变量‘Cabin’的缺失值过多，不便于处理，便将此列直接删除。在观察变量‘Ticket’时，发现此列数据的形式并非全部为数值型，同时也不便于转化为数值型，便将此列直接删除。

之后再观察变量‘Sex’和‘Embarked’两列，发现数据类型均为字符型，为了在后面训练模型时便于处理，便将字符型的数据转换为数字型的数据，其中在变量‘Sex’中，1 代表‘Male’，2 代表‘Female’。在变量‘Embarked’中，1 代表‘S’，2 代表‘C’，3 代表‘Q’。

然后再在所有的数据集中寻找缺失值，发现变量‘Embarked’、变量‘Age’和变量‘Fare’之中均存在缺失值。对于变量‘Embarked’，我们先找到数据的分布情况，发现数值‘1’出现的次数最多，即 1 为整列数据的众数，因此直接用众数‘1’来填补数据列中能够的缺失值。对于变量‘Age’和‘Fare’，我们找到数据列的平均值，并用平均值对数据列的缺失值进行填补。

变量分析

在完成数据预处理之后,剩余的变量为‘PassengerId’、‘Pclass’、‘Sex’、‘Sibsp’、‘Parch’、‘Fare’、‘Embarked’以及要预测的变量‘Survived’。对变量相关性分析时，首先计算出各个变量之间的相关系数，观察变量之间的相关性，画出热力图。

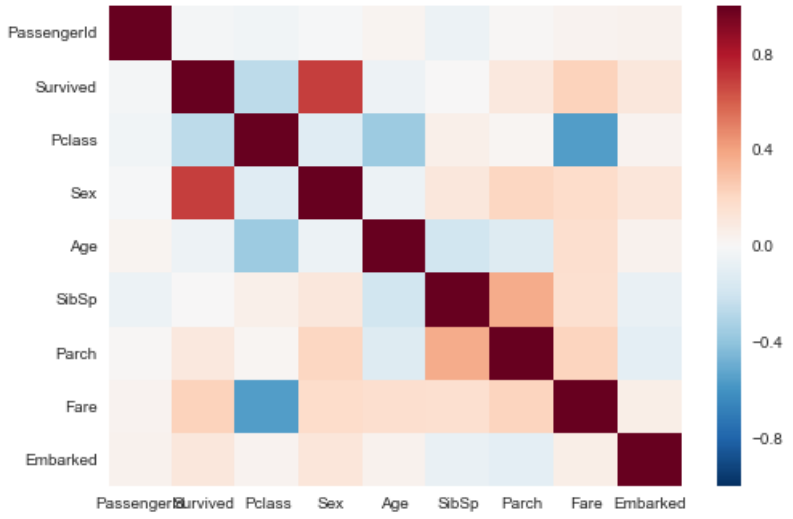

图 2: 变量相关性热力图

由图中各个变量之间的相关性可以看到，变量‘Survived’与‘Sibsp’之间的相关性十分的小，因此这里将变量‘Sibsp’这一列直接删除，变量‘Survived’与‘Sex’之间的相关性十分的大，达到 0.7，变量‘Survived’与变量‘Pclass’之间则呈现负相关。

最终在训练模型中要用到的变量为‘PassengerId’、‘Pclass’、‘Sex’、‘Parch’、‘Fare’、‘Embarked’以及要预测的变量‘Survived’。然后我们分别来分析变量与‘Survived’之间的关系。

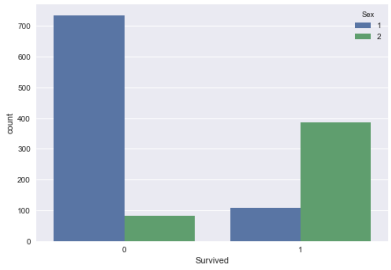

图 3: 变量‘Sex’分布图

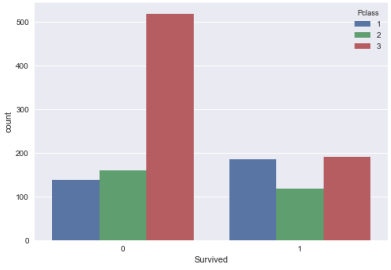

图 4: 变量‘Pclass’分布图

首先由图中变量‘Sex’的分布情况可以看到，获救的人群中大多数为女性，男性则占少数。在未能获救的人群中，男性占多数，而女性数量相对较少。这也说明了在泰坦尼克事件中女性乘客有更大的几率获救。

由图中变量‘Pclass’的分布情况可以看到，在所有乘客之中，选择三等舱的人占大多数，但这并没有使得获救的人群中选择三等舱的人居多，相反在未能获救的人群中，三等舱的人居多，而在获救的人群中，选择不同舱位的人数分布大致均衡。这也说明了在泰坦尼克事件中选择三等舱位的乘客有更大的几率不能获救。

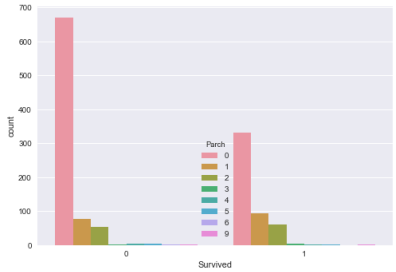

图 5: 变量‘Parch’分布图

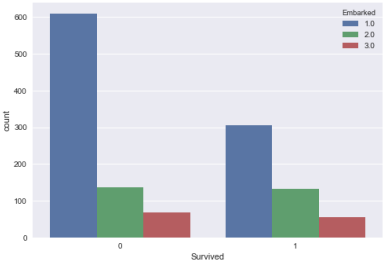

图 6: 变量‘Embarked’分布图

由图中变量‘Parch’的分布可以看出，大部分的乘客都是一人旅行，并没有携带家属，而且在获救人群与未获救人群中，一人旅行的乘客都占大多数，不过一人旅行的乘客有将近 $\frac{2}{3}$ 未能获救，而其他借贷家属的乘客有

50% 的几率可以获救。这也说明了泰坦尼克事件中，携带家属的乘客有更大的几率获救。

同时观察变量‘Embarked’的分布，我们也可以发现类似的规律，乘客之中大部分人会前往 Cherbourg，其他人会前往 Queenstown 和 Southampton。而在前往 Cherbourg 的人群中，只有 $\frac{1}{3}$ 的人获救，在前往其他地方的人群中，有 50% 的人获救。说明前往 Queenstown 和 Southampton 的乘客有更大的几率获救。

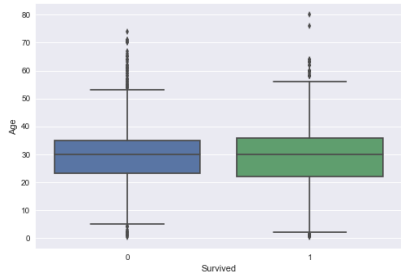

图 7: 变量‘Age’分布图

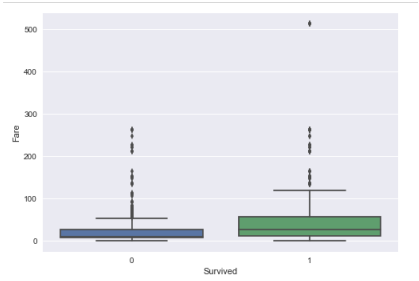

图 8: 变量‘Fare’分布图

观察图中变量‘Age’的分布，可以发现获救的人群与未能获救的人群相比，获救人群数据的上四分位数更大，下四分位数更小，说明了年龄较小或者年龄较大的乘客有更大的几率会获救。

观察图中变量‘Fare’的分布图，可以发现，获救的人群大多为支付的票价较高的人，而相对来说，未能获救的人群大多支付的票价较低。这也说明了在泰坦尼克事件中，支付的票价较高的人有更大的几率可以获救。

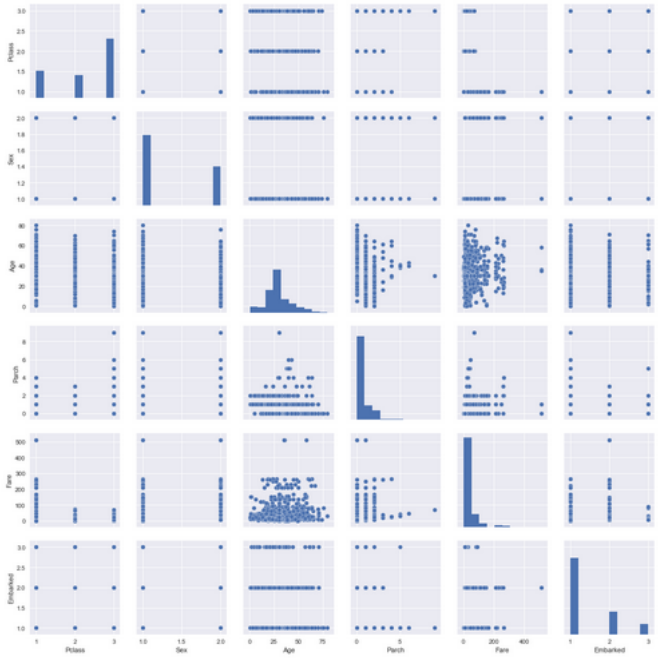

图 9: 变量分布点对图

最后，可以通过数据的点对图更清晰地观察变量数据的分布情况。

模型建立

在建立模型，求解问题时，将数据集中 30% 的数据分为测试集，70% 的数据为训练集。同时采用 KNN 算法、支持向量机、朴素贝叶斯、决策树、逻辑回归和随机森林算法来对数据进行训练，以求从多种不同的模型中获得一个准确度最高，效果最好的模型。

KNN 算法

首先，使用 KNN 算法建立模型，训练数据。在选取 K 值时，将 K 值 1,5,8,9,11,15,20 都进行一遍测试，得到的结果如下表所示。

表 1: KNN 算法准确率

K 值	1	6	9	15	21	30
准确率	0.55	0.60	0.61	0.61	0.61	0.60

由图中结果可以看出 KNN 算法的准确度大都在 60% 左右，普遍偏低。而且从模型计算结果的召回率来看，召回率分布普遍不平衡，对于未能获救人群的召回率较高，普遍为 90%，而对于获救人群的召回率较低，因此导致模型整体的准确率低。同时也可能由于 KNN 算法自身比较简单，计算效率不高，特别是在特征维度大时更是如此，最终模型的准确率不高。

支持向量机

在使用支持向量机模型时，选用 SVC 来进行模型构建，同时对 SVC 的两个主要参数核函数参数以及约束惩罚参数的不同值都进行测试。取约束惩罚参数 C=1，对于不同的核函数‘linear’、‘rbf’、‘poly’、‘sigmoid’得到的结果如下所示。

表 2: 支持向量机 (C=1)

核函数	linear	rbf	poly	sigmoid
准确率	0.87	0.57	0.87	0.87

取核函数为‘linear’，对于不同的约束惩罚参数 C 值，得到的模型结果如下。

表 3: 支持向量机 (kernel=‘linear’)

C 值	0.01	0.1	1	10	100
准确率	0.88	0.88	0.87	0.86	0.85

由表中的准确率结果可以看到，支持向量机模型的效果比 KNN 算法有明显的提高，准确率都在 87% 左右。当约束惩罚参数 C=1 时，对于不同

的核函数参数，可以看到模型结果的准确率并没有太大的变化，同时根据模型结果的召回率也可以看出不同核函数下，召回率的变化也十分小。而且不同核函数下的混淆矩阵也大致相同，模型正确预测的个数变化不大，说明在此数据集中核函数参数对模型结果并不会产生太大的影响。

当核函数为‘linear’时，对于不同的约束惩罚参数 C，当 C 值越大时，惩罚越大，支持向量机的决策边界越窄。在此模型中取 C=0.01,0.1,1,10,100，可以看到模型的准确率在 87% 左右，而且 C 值越大，准确率越低。当核函数为‘linear’，C=0.01 时，模型的准确率达到 0.88，结果最好。观察结果的召回率也可以看到，召回率全部都在 80% 以上，说明支持向量机模型对于求解问题会产生较好的效果。

朴素贝叶斯

在利用朴素贝叶斯模型来求解问题时，利用模型对训练集进行训练，并对测试集进行预测，最终得到的准确率为 64%，对于获救人群和未能获救人群的召回率也只有 49% 和 75%。由此可以看出朴素贝叶斯模型在求解此问题时并不能得到很好的预测效果，可能与模型本身需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。

决策树

决策树是一种基于树形结构的算法，算法本身简单直观，易于理解，具有较高的分类精确度。在利用决策树来解决此问题时，针对不同的叶子数量，对模型分别进行训练及预测。最终得到当叶子数量为 8 时，模型的准确率为 88%，对于获救人群和未能获救人群的召回率分别为 80% 和 94%。由此可以看出决策树模型可以较好地解决此问题，并能够得到较高的准确度。

随机森林

随机森林是在决策树模型的基础上，通过使用随机的方式从样本中抽取样本、选择特征、训练模型，构建很多各不相同的决策树，形成森林。随机森林相对于决策树，在模型求解上会有更好的精确度。在用随机森林求解此问题时，选取不同的决策树的棵数，分别对模型进行训练，最终得到当树木为 100 棵时，模型达到一个较好的效果，准确率为 91%。模型对于获救人群和未能获救人群的召回率分别为 84% 和 96%。由此可以看出随机森林模型能够很好地解决这一问题。

```
             precision    recall  f1-score   support

          0       0.89      0.96      0.92       222
          1       0.94      0.84      0.89       171

avg / total       0.91      0.91      0.91       393

[[213   9]
 [ 27 144]]
0.908396946565
```

图 10: 随机森林结果

在建立不同的模型，并对模型进行训练和预测之后，得到随机森林模型在解决泰坦尼克事件分类问题上可以达到很高的精确度，模型准确率为 91%。模型对于获救人群和未能获救人群的召回率分别为 84% 和 96%。因此最终便会选用随机森林模型来求解这一问题。