**旧金山犯罪种类预测**

**狗的品种分类**

**SY1706216 王晓伟**

北京航空航天大学 计算机学院

15895903730@163.com

【摘要】犯罪根据类型可以分为偷窃，抢劫，杀人等。不同类型的犯罪需要的处理方式和警察力量是不一样的。犯罪类型与该犯罪发生的时间，区域，地点有着很密切的关系，因此，可以通过犯罪的时间，地点相关的信息对犯罪种类进行预测，从而为警方处理这些事件提供依据。本文中，我们通过在2003年到2015年之间的一些犯罪信息的数据集上建立相关的机器学习模型对犯罪种类进行预测，并对结果进行评价和分析。本次用到的主要模型有神经网络，朴素贝叶斯，逻辑回归，决策树和随机森林以及混合模型。

【关键词】犯罪种类预测，神经网络，随机森林，决策树，朴素贝叶斯，逻辑回归，混合模型。

1 引言

通过犯罪发生的时间，地点信息对犯罪种类作出预测是一个多分类问题。在本文中，我们首先对数据集的相关特征进行分析处理，然后分别选择神经网络，朴素贝叶斯，逻辑回归，决策树和随机森林建立模型并评估，再进行模型融合，最后进行总结和评估。本文采用的评价函数是multi class log loss。

本文主要的结构安排如下：在第二节我们介绍对特征相关分析和处理的工作。在第三节我们介绍相关的模型，第四节我们介绍实验结果和进行相关的分析。最后我们对论文进行总结和分析。

2 数据特征工程

2.1 基本特征分析

本文采用的数据集来自Kaggle[1]。数据集包括训练集和测试集两部分。训练集有878049行数据，测试集有884262条数据。训练集有9个属性，分别是

* Dates ：犯罪发生的时间
* Category：犯罪发生的类型
* Descript：犯罪相关的描述
* DayOfWeek：犯罪发生在第几周
* PdDistrict：犯罪发生的区域
* Resolution：犯罪的解决方法
* Address：犯罪发生的地点
* X:犯罪发生地点的经度
* Y:犯罪发生地点的维度

其中，Descript和Resolution仅仅在训练集里面才有，所以我们不考虑这两种属性。首先，我们对训练集的犯罪类型进行统计分析，可以得到图1.从图中我们可以看出，犯罪种类总共有39种，其中，犯罪次数从大到小排在前五名的分别是偷窃，进攻，非刑事犯罪，攻击和毒品。其余的发生的次数比较小。我们接着分析了犯罪发生和经纬度之间的关系，我们通过画图，得到图2。从图中我们可以看出，绝大部分的犯罪经度在-122.2-122.6之间,绝大部分犯罪位于36.38-40之间，而还有很少的点在纬度超过90度的地方，我们认为这是噪声点，应该数据清洗过程中从数据文件中删除。我们接着对每个区域发生的案件个数进行分析，如图3,我们可以看到，在Southern发生的犯罪案件最多，而在Richmond发生的犯罪案件最少。

2.2 数据清洗

根据数据分析中的相关信息，我们将纬度大于40度的点从训练集中删除，生成新的训练集。

2.3 特征选择

针对每一个模型的特点，我们选择生成四个数据集。四个数据集信息分别如下：

* 数据集1

数据集1中，我们选择日期中的小时(hour)，犯罪发生的区域(Pdstrict),以及第几周（DayOfWeek），对这些特征进行二值化，同时将犯罪类型(Catagory)转换成数字标签，生成数据集1

* 数据集2

在数据集1的基础上，我们加入Address这个特征，

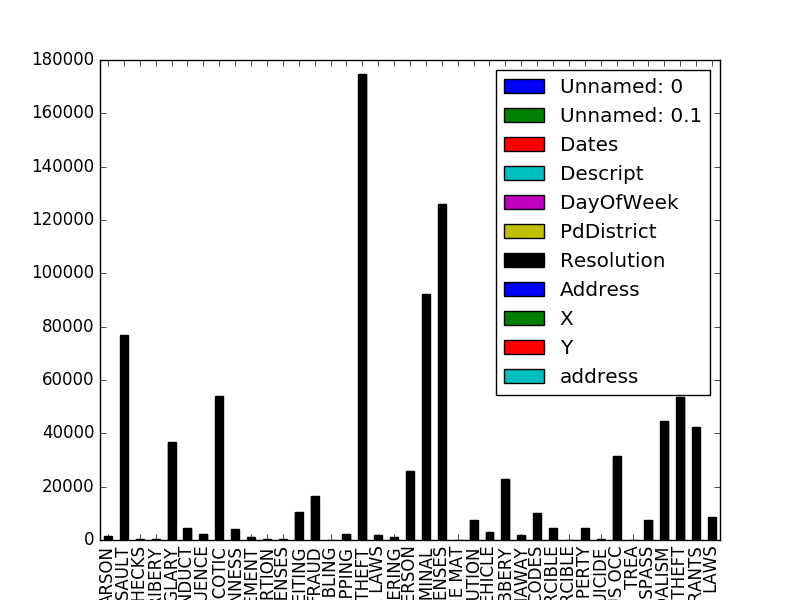


图1 犯罪类型次数统计

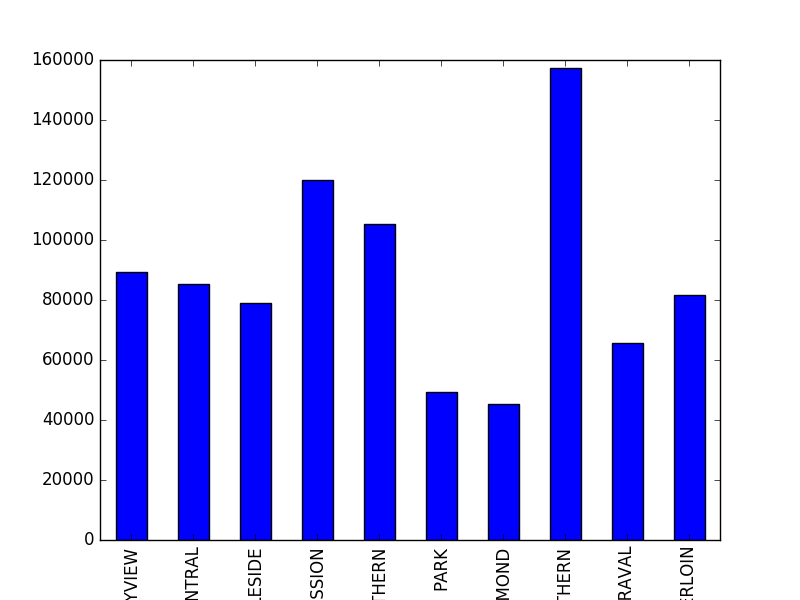


图3 每个区域犯罪次数统计

* 数据集3

在数据集1的基础上，我们加入X和Y的信息，生成数据集3

* 数据集4

在数据集3的基础上，我们加入将Adress换成每一个Adress犯罪发生的次数，并生成数据集4.

* 数据集5

在数据集3的基础上，我们将地址的街道信息拆分出来，形成StreetNo这个特征，然后根据小时确定是否是白天还是晚上，再加一个特征。加入Adress这个特征，生成特征集5。

3 模型

我们分别使用朴素贝叶斯，神经网络，决策树和随机森林，逻辑回归以及混合模型来对数据进行预测，下

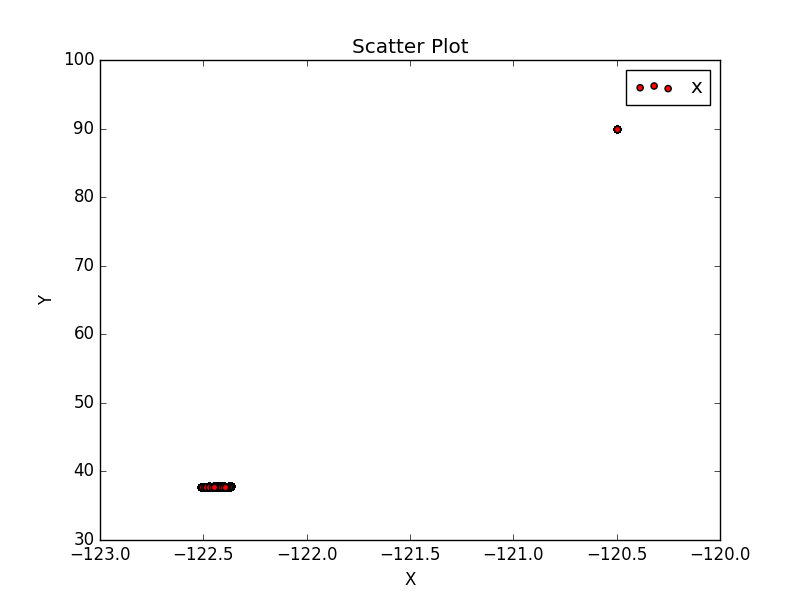


图2 犯罪经纬度分析

面我们分别讲解每一个模型的基本原理。

3.1 朴素贝叶斯

朴素贝叶斯的基本思想是用先验概率来预测后验概率。给定一个D维向量X,预测X属于的类别Y。由朴素贝叶斯公式可得

(2)

朴素贝叶斯假设变量之间独立。因为X和Y是连续值，而且有较强的相关性，所以我们只在数据集1,2上对朴素贝叶斯进行建模，将转换好的Catogory作为类别标记，用sklearn进行建模训练并预测结果。

3.2 神经网络

神经网络是一种常用的分类器。神经网络通过模拟人脑的神经元的结构，建立多层的神经网络，然后训练模型，并对相关数据进行预测。神经网络由输入层，隐层和输出层三部分组成。其基本的模型图如图[3]所示，神经网络通过BP反向传播算法来优化模型，使模型达到最佳分类准确率。

在本次实验中，我们分别用数据集1.3.4来训练生成模型并对数据进行预测，从而比较数据集的不同对神经网络准确率的影响。同时，我们还比较在同一数据集上，不同得参数对神经网络预测准确性的影响并生成最后的结果。

3.3 决策树

决策树是一种树形结构，由节点和有向边组成。内部节点表示一个属性，叶节点表示一个类别。分支代表一个类别的具体值。决策树通过自顶向下的方法构建，每次选择可以使信息增益最大的属性进行分类，最终生成相应的树形结构。预测时，则根据相应的规则进行依次分类直到叶节点。

本次实验中，我们在数据集1上对决策树进行测试。

3.3 随机森林

随机森林是一种集成学习方法。随机森林以决策树作为基学习器。对于基决策树的每个结点，先从该结点的d个属性值中随机选择一个包含k个属性的子集，然后再从这些子集中找一个最优子集进行划分，最后综合所有决策树的结果生成随机森林。

本次实验中，我们在数据集2和5上对随机森林进行训练并输出预测结果。

3.4混合模型

我们在上述基本模型的基础上，对每一个模型预测结果进行线性组合，然后生成混合模型，并对混合模型的结果进行测试。混合模型用并行的生成方法，对每一个基本模型生成的结果进行线性组合，然后生成最后预测结果。

实验中我们在通过不同的数据集训练生成基模型，然后分别预测，对最后结果根据基模型准确率进行线性组合。并生成最终结果。

4 实验结果

我们将上述模型进行训练并在kaggle数据集上进行测试。得出如下结果。以下出朴素贝叶斯外，别的准确率都为kaggle上测试结果。

4.1 朴素贝叶斯实验结果

朴素贝叶斯的模型结果如表1。从表中我们可以看出来，我们将训练集分为按照0.6和0.4的比例分为训练集和测试集，分别在测试集和最终kaggle上进行评估，得到结果如表1。

表1 朴素贝叶斯模型训练集和测试集muti-log损失

|  |  |  |
| --- | --- | --- |
| 特征集 | 训练集log损失 | 测试集log损失 |
| 1 | 2.58135 | 2.58375 |
| 2 | 2.56377 | 2.56148 |

从表中我们可以看出，朴素贝叶斯的muti-log损失在训练集和测试集上非常接近，这符合我们的常识，朴素贝叶斯是一个稳定的模型，受样本影响比较小。

4.2 神经网络实验结果

我们将分别在数据集1,3,4上训练神经网络模型，并对每一个神经网络模型调节参数并返回相应的结果，得到表2。和表3。

表2 神经网络模型在数据集1和3上相关结果

|  |  |  |
| --- | --- | --- |
| 隐层参数 | 特征集1 | 特征集3 |
| 40,40 | 2.59325 | 2.57377 |
| 100,100 | 2.60597 | 2.57787 |
| 40,40,40 | **2.57524** |  |

表3 神经网络模型在数据集4

|  |  |
| --- | --- |
| 隐层参数 | 特征集4 |
| 40,40 | 2.50626 |
| 40,40,40 | 2.53265 |
| 30,30 | 2.51716 |
| 80,80 | 2.53131 |

可以看到，神经网络的预测结果随着特征的丰富不断增强，具有较强的学习能力。但是随着神经网络隐层的增加，神经网络在训练集上的muti-log损失越来越小，在测试集上的muti-log损失越来越大。所以非常容易过拟合。

4.3 决策树实验结果

我们在数据集1上建立决策树的模型并进行测试，得到最终的结果如表4.

表4 决策树模型结果

|  |  |
| --- | --- |
| 决策树参数 | 特征集1 |
| 准侧函数:gini 最大深度:无 最小样本数:2 | 4.17 |

4.4 随机森林实验结果

我们在数据集3和5上对随机森林进行测试，最终得出的结果如表5和表6。可以看出，xgboost具有比sklearn更好的训练效果。

表5 随机森林模型结果

|  |  |
| --- | --- |
| 随机森林参数 | 特征集3 |
| 1. 50棵数，最大深度20， | 2.57947 |
| 50棵数，最大深度15, | 2.56802 |
| 100棵数，最大深度20, | 2.65188 |
| 100颗树，最大深度15， | 2.57446 |

表6 随机森林模型结果

|  |  |
| --- | --- |
| 随机森林参数 | 特征集5 |
| 1024棵数，最大深度15， | 2.37226 |

4.5 混合模型实验结果

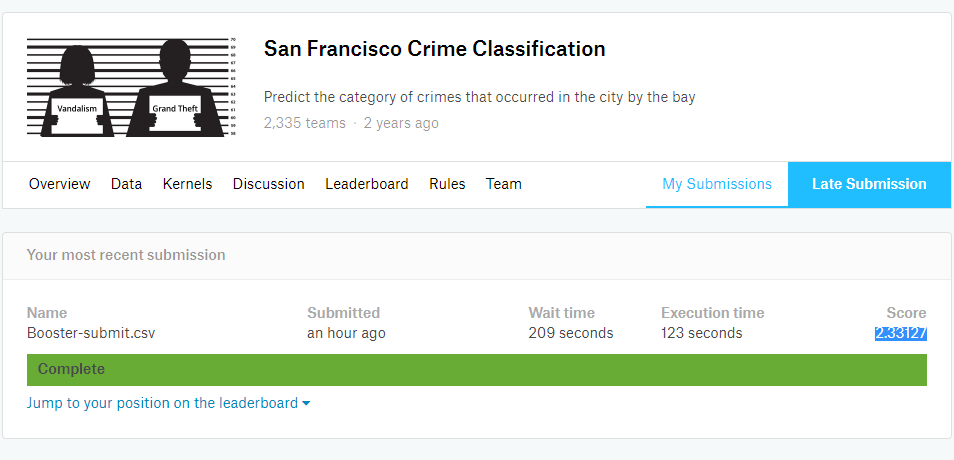
混合模型采用并行混合的方法，将每一个模型输出的预测结果根据模型本身的准确性相加，最终得出结果。混合模型模型间数据集独立。本次混合模型主要做了两组实验，一组是贝叶斯模型和神经网络的混合模型，一组是朴素贝叶斯，神经网络和随机森林的混合模型。训练结果如表7可以看出来，混合模型的准确率优于单个模型的准确率。

表7 随机森林模型结果

|  |  |  |
| --- | --- | --- |
| 基分类器极其准确率 | 集成策略 | 混合模型准确率 |
| 朴素贝叶斯h1:准确率2.5604  神经网络h2:准确率2.64 | F(x) = 0.6\*h1(x)+0.4\*h2(x) | 2.55735 |
| 朴素贝叶斯h1：2.5604，随机森林h2:2.5887 神经网络h3:2.60 | F(x) = 0.6\*h1(x) + 0.2\*h2(x) + 0.2\*h3(x) | 2.54329 |

4 kaggle结果展示

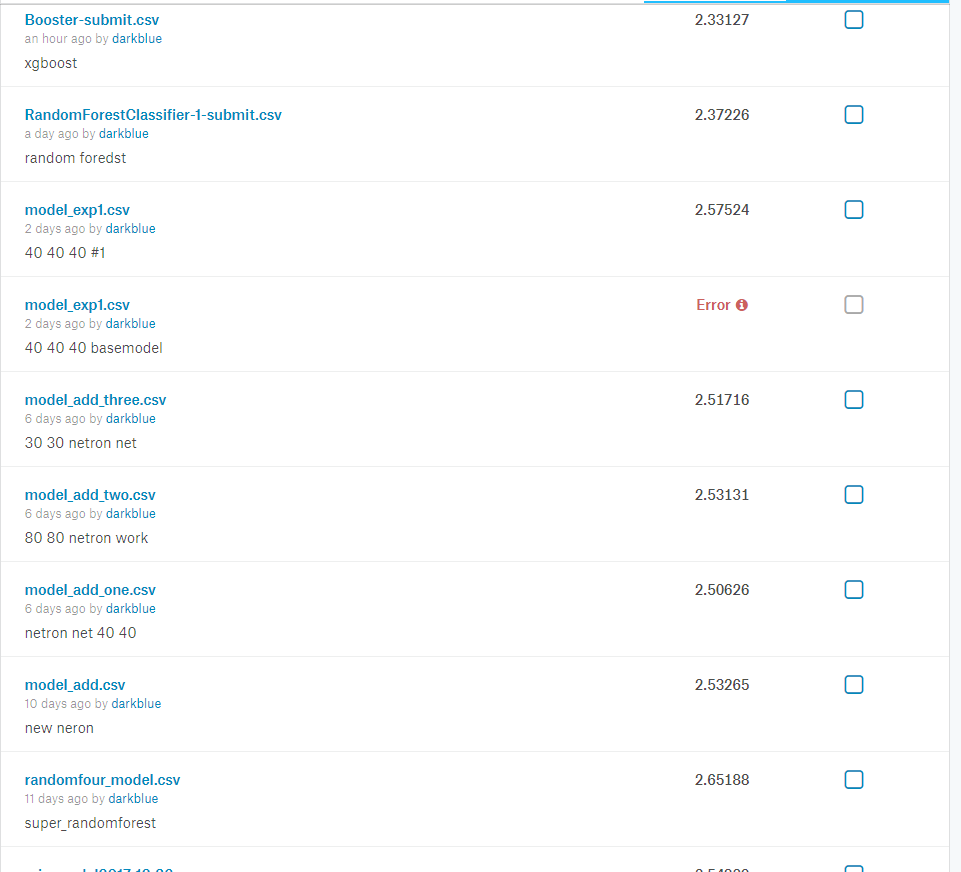
Kaggle的用户名为darkblue,本次在kaggle上最后的分数是2.33127。个人最好结果如图8，为2.33127。个人测试结果如图片9所示。



**图片8个人最好分数**

5个人总结

本次个人作业自己收获了很多。在特征处理方面，如何处理离散化的特征，如何挖掘它们的信息，以及不同数据集对应的分类效果的变化，自己都有了很深的理解。同时，自己还测试了串行和并行的集成学习模型，充分



**图片9 kaggle提交结果**

了解集成学习对分类准确率的提高作用。最后，自己先后实验了朴素贝叶斯，决策树，神经网络和随机森林，以及xgboost的boosting分类器，得出相关参数和准确率的变化关系，本次个人项目有很大收获。

References (参考文献)

[1]周志华. 机器学习[M]. Qing hua da xue chu ban she, 2016.

[2] Junbo Ke.Computer Science and Engineering University ofCalifornia,San Diegojuke@eng.ucsd.eduSan Francisco Crime Classification.

**狗的品种分类**

**小组作业**

**SY1706216 王晓伟**

**身份:组长**

1. 大作业分工

本次大作业我是组长，我主要负责的内容是前期关于相关方法的收集和评估。中期自己选择几个较优的方法并分配任务。自己在期间主要负责迁移学习模型的训练和评估以及优化。后期综合每个模型的特点进行交流和总结。

2. 狗的分类任务概述

图像识别是机器学习中的一个重要的领域。本次小组大作业我们小组做的是狗的品种分类，输入为狗的图像，模型输出为狗的种类。本次任务是一个多分类的任务，评价标准是muti-logloss。

3. 数据特征工程

我们将图片预处理成tensorflow.record文件方便进行读取，并将来别处理成多维0,1向量。我们对每种狗的种类个数做了统计，如图1所示。

4. 迁移学习理论

迁移学习核心思想是用已经学到的知识来学习新的知识，关键是找到原有知识和新的知识的相关性，通常是把一个已经训练好的模型参数用在训练另一个模型上。迁移学习可以避免我们任何时候都要从零开始去训练一个模型

而且在训练数据量不足的时候，迁移模型可以很好的利用之前模型的相关数据，从而解决数据量不同的问题。在本次任务中，因为狗分类属于图像分类的领域。在本次的例子中，狗的种类有120多种，而图片只有20k张，所以数据量不够。解决这类问题有两种方法，一种是增加数据量，另外一种是迁移学习，使用之前已经训练好的图像分类的模型，将这个模型迁移到本次数据集上进行训练。本次大作业采用迁移学习的方法来进行模型训练和评估。

5. 模型建立

我们首先下载谷歌一个已经训练好的模型，然后在这个模型的头部加上三层全连接的网络，模型的示意图如2。首先图像输入给先验模型，然后在先验模型后面接一个连接成，连接层的结点个数和先验层最后一层的结点个数相同，结点个数为2048，然后后面加一层自定义的网络隐层，参数可以调节。最后一层为输出结点，数目跟狗的种类数一样为120个节点。模型的参数总共有4个。隐层的网络节点数，并用数据对这个新的模型进行训练。为了节省时间，本次直接用全部训练集训练模型，然后对kaggle数据集进行测试并输出最终结果。针对不同的参数进行训练并输出结果。可以看得出来，迁移学习的效果最佳参数在隐层结点是2048个结点。并且迭代次数越多，muti-log损失越少。

自定义网络层

自定义网络层

连接层

先验模型

图像











图 2 迁移学习模型

6. 实验结果

表1 迁移学习模型参数以及对应结果

|  |  |
| --- | --- |
| 模型参数 | Kaggle muti-log损失 |
| 隐层网络结点=1024，批大小=64,迭代次数=5000,学习率=0.001。 | 0.10037 |
| 隐层网络结点=2048，批大小=64,迭代次数=5000,学习率=0.001。 | 0.07617 |
| 隐层网络结点=2048，批大小=64,迭代次数=6000,学习率=0.001。 | 0.03750 |

7. kaggle具体测试结果

我的Kaggle用户名为darkblue，kaggle数据集如图所示，最好的分数是0.03左右，排名在第50/669名。大概在前10%左右。结果如图3,4所示。

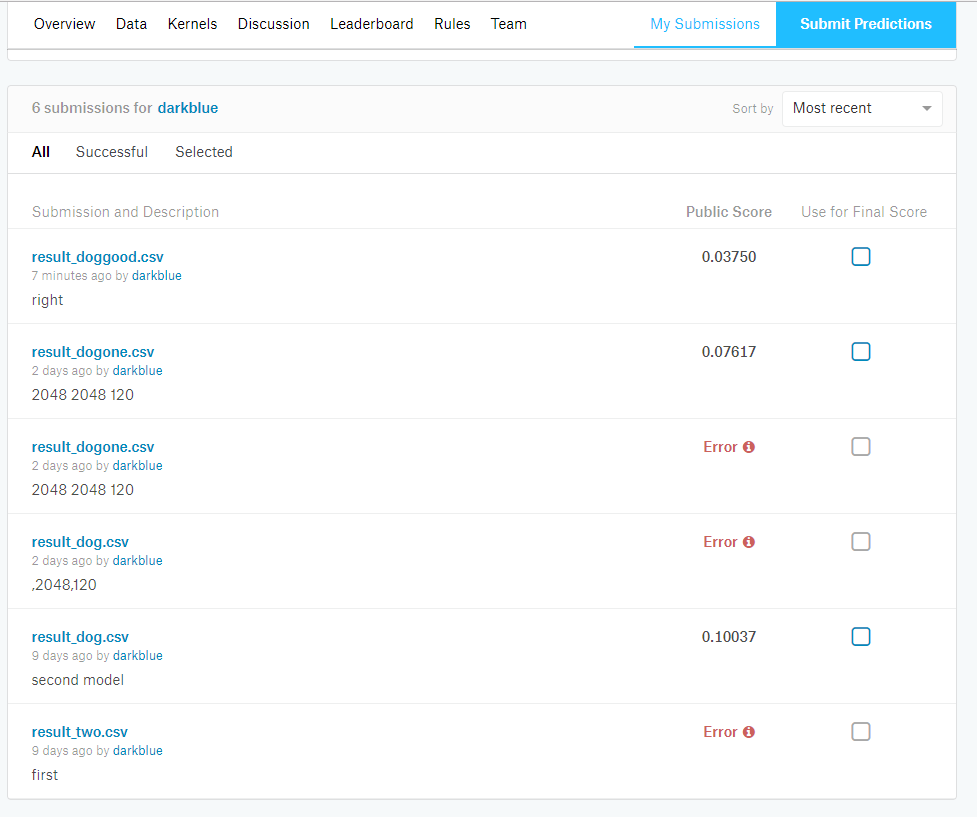


图3 kaggle个人提交记录

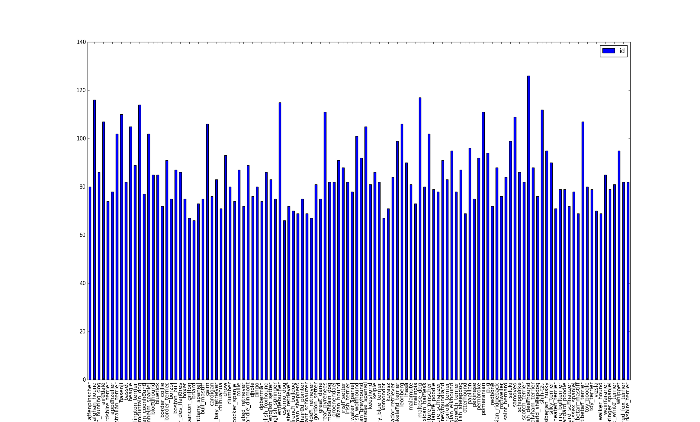


图1狗的种类统计

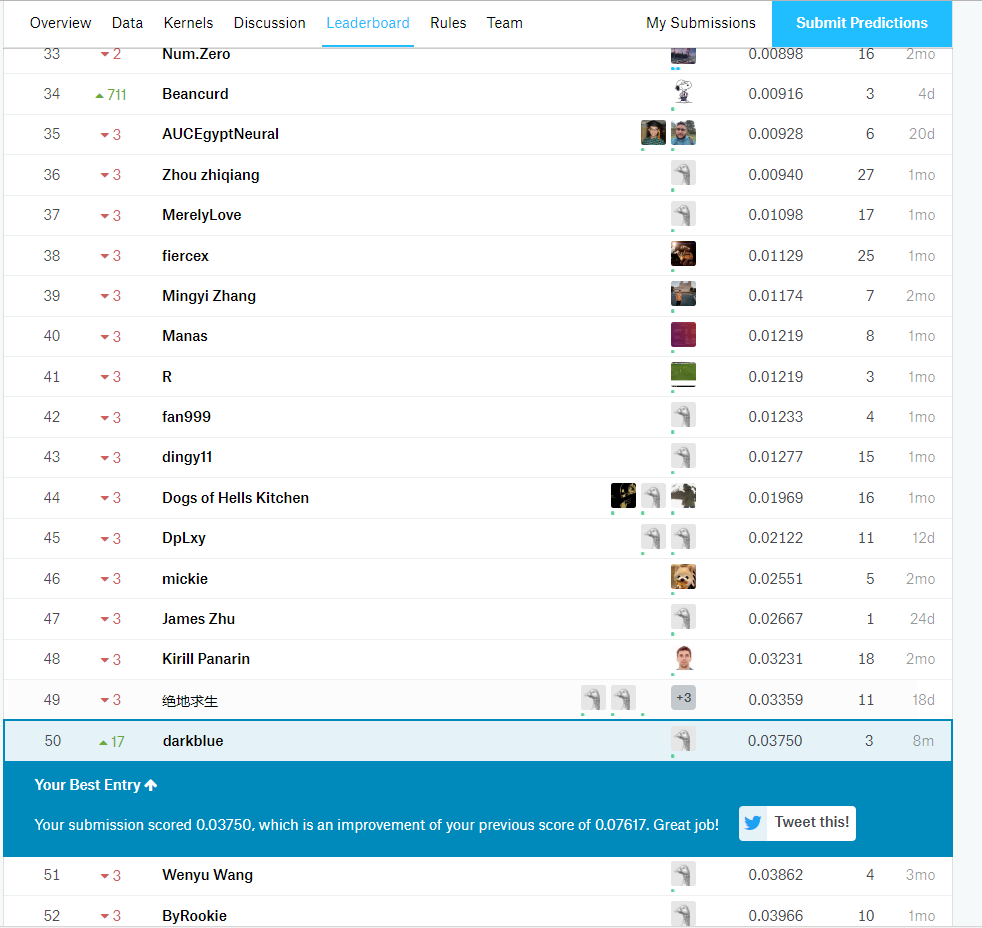


图4 kaggle公共排行榜

9. 个人总结

本次大作业自己学到了很多东西。首先是计算机视觉方面关于图像处理的一般方法，图像的分析方法。还有tensorflow框架的用法。理论方面，自己学习了迁移学习的相关理论和知识。迁移学习可以充分运用之前的一些工作和知识，有着很好的继承性。可以避免从0开始花费巨大的时间去训练模型，具有非常多的优点。唯一遗憾的是自己的机器太差，没办法训练深层的迁移学习模型，不然应该可以有更好的名次和结果。

References (参考文献)

[1]周志华. 机器学习[M]. Qing hua da xue chu ban she, 2016.