欧洲足球数据分析

该项目代码见 https://github.com/rucnyz/soccer ，代码按照下述说明完成，第一次运行（未抽取特征并保存）需要 71 分钟左右。

1 数据集

欧洲足球数据集有着丰富的球队、球员以及比赛数据。具体来说，包括从 2008 年到 2016 年超过 25000 次比赛数据、超过 10000 名球员的数据来自 11 个欧洲国家自己的联赛，其中球员和球队的能力数据来源于 EA 游戏 FIFA 的内容。而且每场比赛还包括 10 个博彩网站的赔率数据。

2 初步思路

2.1 数据集分析

面对非常丰富的数据以及各种各样可选择的任务，首先对数据集进行梳理并明确目标。

总的来说，我们认为和本课程关联最大且最有意义的研究内容包括以下两个部分：（*待补充*）

描述分析	预测任务
用雷达图描述球员能力， 研究球场不同位置雷达图的区别	对比赛结果进行预测 （多分类任务）

原因：该数据集包含大量 FIFA 球员数据以及丰富的球员特征，应当能够得到良好的效果；同时包含 2008-2016 年 25797 条比赛记录，进行分类任务的数据量足够。

而数据集的核心内容可以分为三个部分（*表格待美化*）：

球员能力	**overall_rating** potential preferred_foot attacking_work_rate defensive_work_rate crossing finishing heading_accuracy short_passing volleys dribbling curve free_kick_accuracy long_passing ball_control acceleration sprint_speed agility reactions balance shot_power jumping stamina strength long_shots aggression interceptions positioning vision penalties marking standing_tackle sliding_tackle gk_diving gk_handling gk_kicking gk_positioning gk_reflexes
球队情况	(buildUpPlaySpeed buildUpPlaySpeedClass buildUpPlayDribbling buildUpPlayDribblingClass buildUpPlayPassing buildUpPlayPassingClass buildUpPlayPositioningClass) (chanceCreationPassing chanceCreationPassingClass chanceCreationCrossing chanceCreationCrossingClass chanceCreationShooting chanceCreationShootingClass chanceCreationPositioningClass) (defencePressure defencePressureClass defenceAggression defenceAggressionClass defenceTeamWidth defenceTeamWidthClass defenceDefenderLineClass)
比赛数据	(**home_player_1~11 away_player_1~11**) (**home_team_goal away_team_goal goal**) (shoton shotoff foulcommit card cross corner possession) (**B365 BW IW ...**)

那么确定了可行任务以及可利用的数据后，接下来就是对该问题进行建模了

2.2 初步研究

分为描述分析和统计建模两个部分。

2.2.1 描述性分析

简单描述

（*待补充*）

2.2.2 统计建模

从上述的数据集分析中可以看出，要预测一场比赛结果，**参赛球员的能力**、**两只球队的历史实力**以及**比赛当时的情况**都是需要纳入考虑的。但是由于特征过多，作为初步的尝试，我们决定尽可能选取较为简单但全面的特征进行研究。

幸运的是，数据集很好的满足了我们的考量，我们的思路如下：

- **球员能力**
  在 Player_Attributes 表中有 FIFA 对球员的整体评分，也就是 overall_rating，然而球员的状态是在不断改变的，只使用一个来表示他过去打的所有比赛未免过于僵硬，好在数据集提供了每个球员 2008-2016 年中多个时间点的状态，我们对每一场比赛都选取了距离该场比赛最近的球员状态作为特征，一共 11*2 个。
- **球队历史实力**
  很遗憾，在 Team_Attributes 表中并没有能够整体代表该球队实力的特征。简单对该球队的球员能力进行加权求和是很不合理的，一方面相较于球员能力的特征，这

样的线性组合属于冗余信息；另一方面简单的加和没有考虑到球员之间的相互作用与配合。

在初步思路中暂时不打算处理过于复杂的情况，因此我们决定将球队过去数场比赛的情况作为当前的整体状态进行考虑（毕竟归根结底一场比赛最重要就是要赢球）。同时再考虑一个该球队所属的联赛。

于是在这里我们生成了多个特征，包括过去 x 场（在代码中采用的是 10 场）分别作为主队和客队的总净胜球数（如果输了就是负的）、过去 x 场分别作为主队和客队的总胜场数和输场数、和这场比赛的对方球队过去 y 场（代码中采用的是 3 场）赢球次数和输球次数、该球队所属的联赛（哑变量）。

- **比赛当时情况**

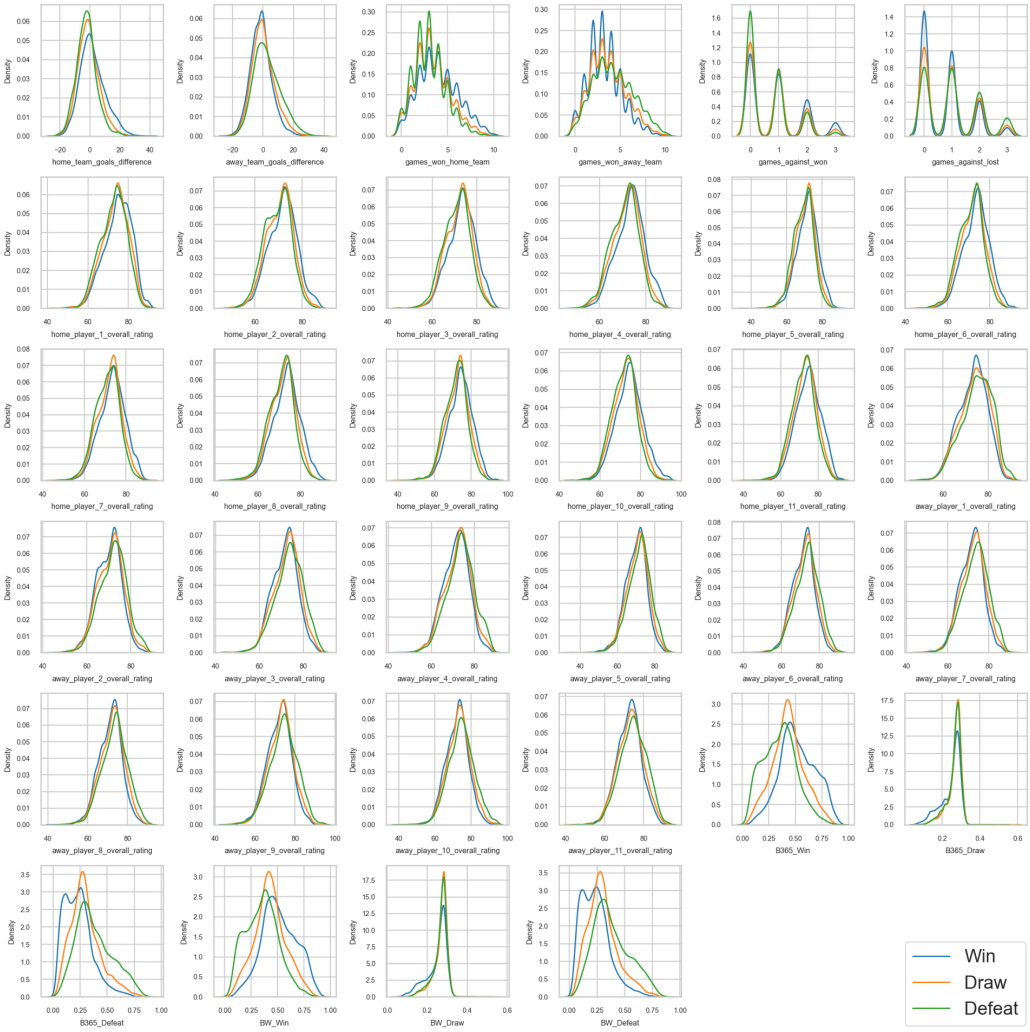
  这个在当前数据是最难考虑的一个地方，因此我们暂时没有将比赛中统计数据进行融合，而是采取了另一个思路。

  我们使用了数据集提供的 10 大博彩网站的赛前赔率数据（为方便起见只用了 Bet365 和 Betway 两个最知名的公司），我们希望赛前的赔率能够包含除开球队、球员数据之外的其他因素，比如赛前的大事件、比赛场地等等因素，从而提高预测胜负关系的效果。

而接下来，则是模型的选用了，作为初步思路我们决定首先选取一些机器学习方法进行研究。同时注意到在特征的选取过程中应当存在不少冗余的情况（赔率数据和前两类数据的关系、球员能力和球队历史胜场情况的关系），我们决定使用一些盲源分离算法首先对数据进行降维处理，再随后再通过机器学习分类器。

那么具体思路如下：

1. 首先将数据集按照上述要求进行特征提取，并随后去除缺失值并进行归一化处理，最终得到 21245 份比赛样本，45 个特征。标签则有 3 类，赢球、平局、输球，比例为 6:3:4。所有特征按照标签进行分类后的密度图见下图。
2. 随后对数据集以 4:1 的比例划分训练集与测试集，而训练集再使用五折交叉验证进行划分。
3. 模型方面的选取分为两个部分，盲源分离算法包括 PCA 和 ICA，分类器包括 GradientBoosting、RandomForest、AdaBoost、NaiveBayes、KNeighbors、LogisticRegression。将完整的模型定义为"先将原始数据通过一个盲源分离算法，再将输出结果传入分类器进行训练"的管道
4. 选择各模型的合理参数范围，将上述所有模型进行网格搜索，使用准确率(Accuracy)作为评价指标，以得到最佳的降维方法和分类器组合。



2.3 初步结果与调整

初次尝试中，我们的配置如下表所示：

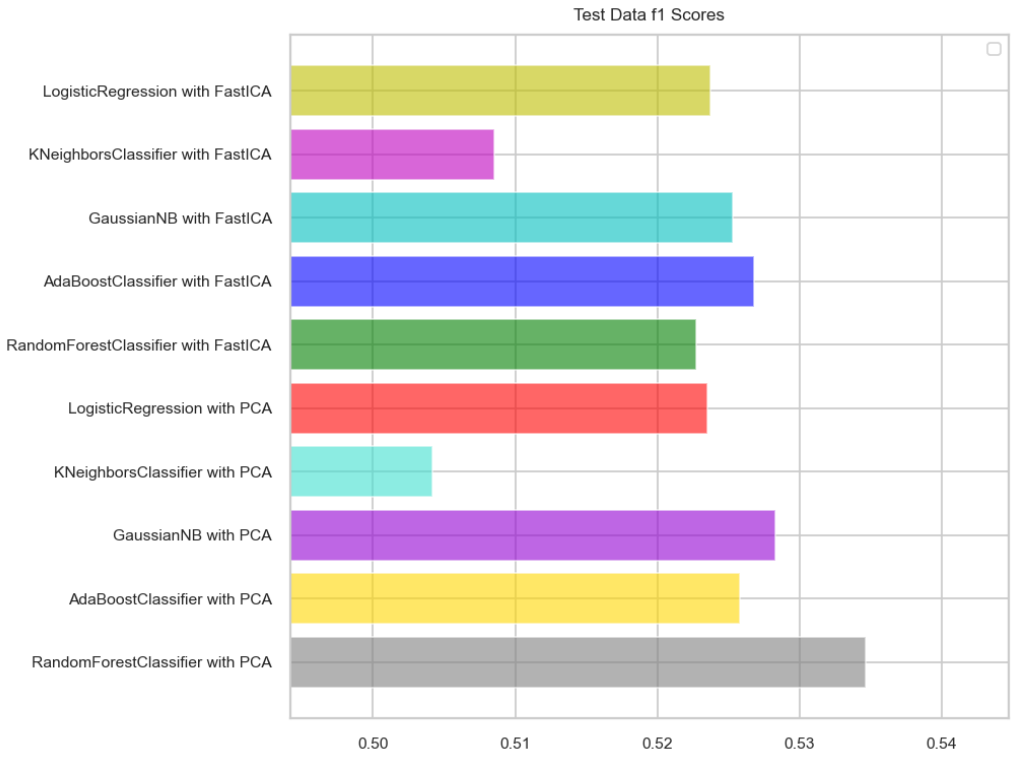
数据集配置		**网格搜索配置**		
数据种类	**比例**	**模型种类**	**配置**	
训练数据	0.8	PCA&ICA	**n_component**	arange(5,46,8)
测试数据	0.2	RandomForest	**max_features**	[auto、log2]
交叉验证配置			**n_estimators**	[50, 100, 200]
打乱次数	**训练验证比例**	AdaBoost	**learning_rate**	linspace(0.5,2,5)
5	4:1		**n_estimators**	[50, 100, 200]
		NaiveBayes	-	
评价指标		KNeighbors	**n_neighbors**	[3, 5, 10]
balanced accuracy	f1 score	LogisticRegression	**C**	logspace(1,1000,5)

考虑到数据不均匀的问题

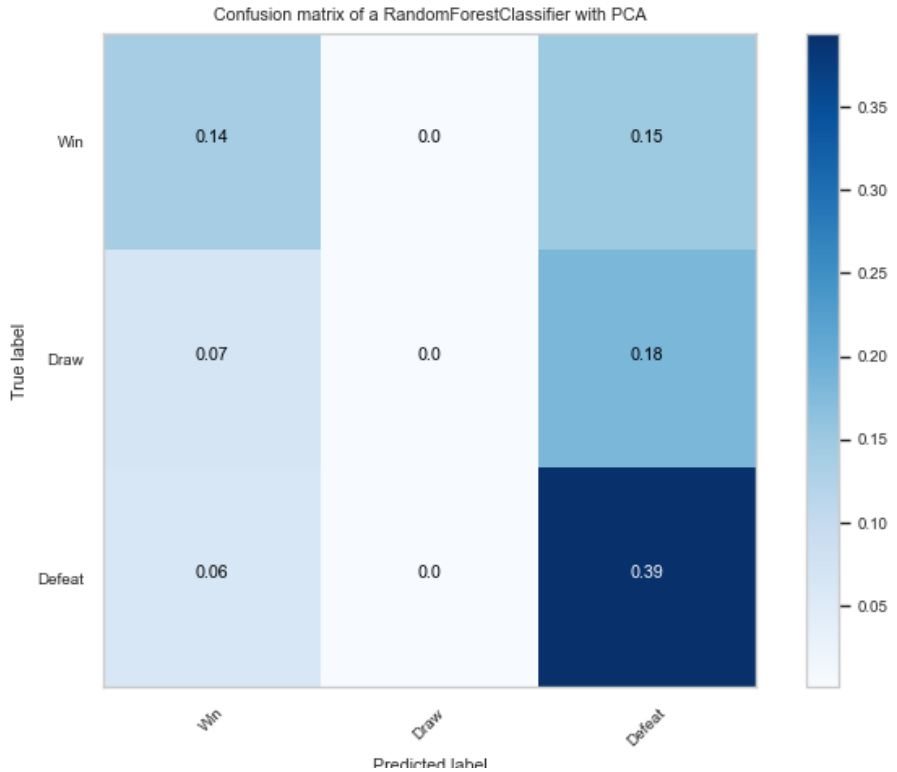
1. 我们采用了分层交叉验证(StratifiedShuffleSplit)，使得交叉验证中的训练集和测试集标签比例适中。

我们使用了 scikit-learn 中自带的 balanced_accuracy_score 针对不平衡的样本使得准确率指标更有意义，而我们的网格搜索使用的评价指标则是 f1_score。

得到的结果如下图所示，结果显示，该情况下使用 PCA 的随机森林算法在该情况下效果最好，但也应当注意到，实际上各个模型之间的差异不大，这样小的差距可能会受到很多随机影响例如随机数种子的设置，我们认为这和抽取的特征过于简单不够丰富有关，这也是我们希望在下一阶段克服的问题。



最好模型的混淆矩阵如下



3 进一步设想

3.1 动机

针对上述得到的信息，我们提出了可以在接下来进行尝试的内容。

描述分析方面

- 将各场比赛中对应位置的球员维度数据取平均值，根据与平均值的接近程度挑选对应球员，这样应当可以挑选出各个位置具有代表性的球员六维模板。
- 每个球队可否通过六维图的方式同样进行描述？

统计建模方面

- 在上述描述分析中，我们采用雷达图对球员个人能力进行了描述，这样的六个维度相比于总体评分应当是一个更加全面以及良好的描述，因为可以针对不同位置描述不同情况。因此我们可以将这样 22*6 个特征纳入模型考虑。
- 一个球员可能可以踢多个位置，在一场比赛中他一定踢到了自己最适合的那个位置吗？如果将球队中球员的位置进行调整，代入上面得出的模型，是否可以得到更大的胜率（这里的对手球队可以采用其余球队球员的数据取平均）
- 球队整体数据考虑过于简单（*待补充*）。
- 这样如果采用更多的特征，应当使用深度学习方法和前述的机器学习方法进行对比。
- 在预测比赛胜负时，明星球员的个人能力堆叠和球队整体的配合水平，哪一个对比赛的走向影响更大呢？

- 我们对球队之间的关系的建模是通过过去数场的净胜球和胜场进行的，过于简单。是否可以采用图神经网络表征球队之间复杂的实力关系呢（比如 A 球队对 B 球队胜率高，对 C 球队胜率低，然而 C 球队对 B 球队胜率高）？
  即使用无向图来对球队之间的胜负关系进行建模，边的权重用两只球队的胜场数除以输场数来表示，边的表征为比赛的情况，节点的表征为球队的实力，这样将预测比赛胜场的问题转化为了链路预测问题。
  但这个思路只能应用于某一个赛季或者将多个赛季分为多个图来表示，和前述内容没有了对比性。