Project Proposal

Richard Yang
Zhoujingpeng Wei
Jiaxin Hu

October 25, 2018

I Introduction

I.1 Overall description

We has collected a dataset containing soccer games from specific leagues recent years, including all kinds of features, which are reflecting the information of teams' performance about attack or defence .

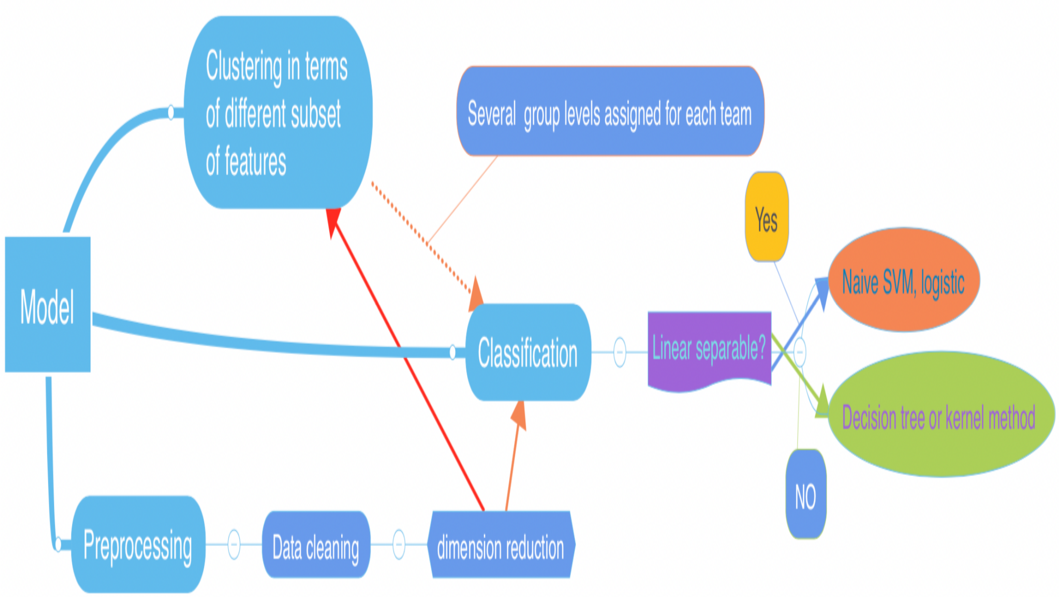
What we want to do is to find the how to choose strategy when confronting different opponents, and what we do is quite **statistical**, put much emphasis on **model explanation**.

I.2 Related work

Following is our procedure:

- Implement clustering: Because our data is just about games, we need to find the style of each team. And we choose to average the games data to find the general performance of each team, but so many features is redundant because of collinearity, so we can just specify some groups by clustering the team's performance according to different feature subsets.
- Analyze the level pattern: we can analyze the clustering result to determine the meaning of the groups: for example, **Chelsea** maybe attributed to a high level group of defence but low level one of dominating the ball.
- Using each kind of group levels above as a numerical feature for each team and their opponents, along with other features about games, we try to fit the game results in terms of these features by some classification algorithm.
- Explain: Try to explain the relationship among the game results, teams levels and games strategies according to decision boundary.
- Evaluate: Test the model on the validation set to do model selection and prove the correctness of our assumed relationship as well, and then we can use that relationship to choose strategies in the future data.

The following graph can make it clear:


Clustering in terms of different subset of features
Several group levels assigned for each team
Yes
Naive SVM, logistic
Model
Classification
Linear separable?
Decision tree or kernel method
NO
Preprocessing
Data cleaning
dimension reduction

Our naive plan

II Motivation

Soccer is the most charming sports worldwide, not just because it's a sport showing good skill, reasonable strategies, but also the hot-blood atmosphere as well as the impressive team culture. And the **World Cup** is the most typical representative, which excites the world every 4 years.

However, in the World Cup of 2018, it seems that the team with high general performance, like good ball possession or pass accuracy, couldn't usually win the game, instead, the team choosing conservative strategies, like counterattack, tends to gain the success, like the final winner–France.

Does that mean the data for general performance is counterintuitive in modern game? or it's just some coincidence? Is choosing the conservative strategy always a good idea? So if we can figure out the true inner logic(which would be extremely tricky), we are able to get the information how to choose strategies to attain ideal result. And we also would have a in-depth understand of in what direction the soccer develops, which would be essential to the professional soccer history. And in terms of data of attack or defence, we can analyze the style of a team and thus can predict its pattern in the future games or provide some useful guidance.

Assume a rich club, like **Manchester City**, hires us to investigate into it, that's because Guardiola takes much emphasis on strategy. Then we want to find a good model to detect such logic, in order to have a good view of how features would influence results of games generally. And we know, every game has much property of randomness, we just model the determined part, and make assumption to omit the random noise.

III Evaluation

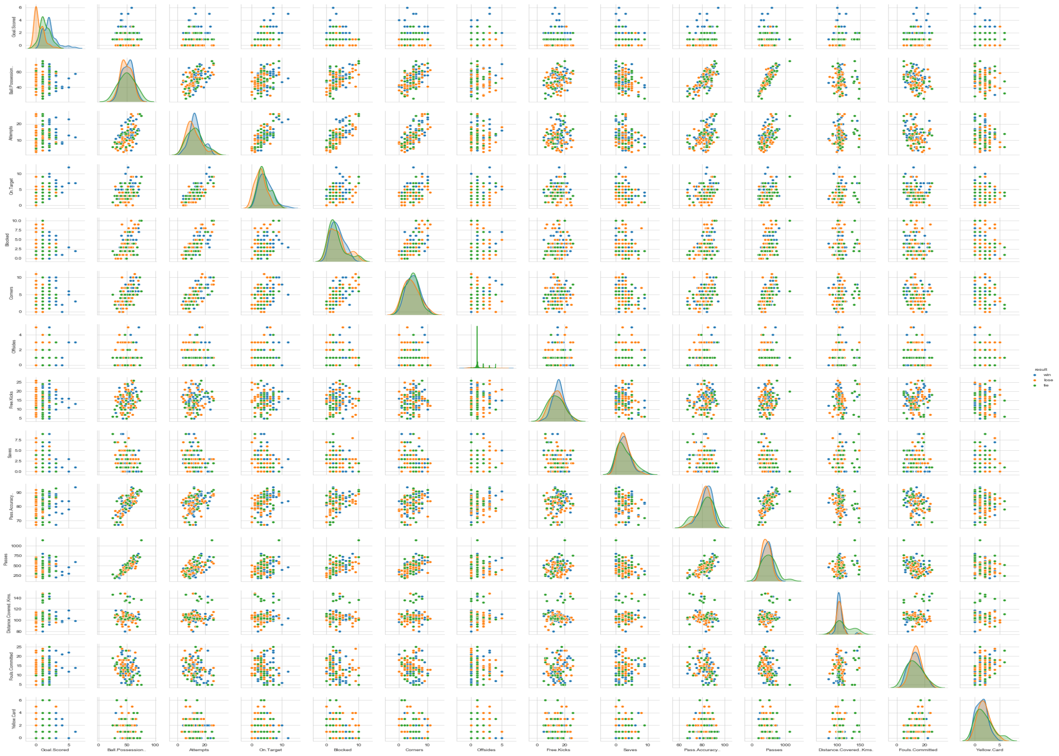
III.1 The expectation of our model

- Our model achieves a good performance in terms of accuracy on the validation set.
- Our model doesn't show an obvious sign of overfitting, such that one leaf node just contains few examples.
- Model can account for all connection between response and features.
- We can have an in-depth view of data, digging all information out, like what can cause more goal, what can cause more lose, and what kind of strategy can cause a better result when confronting different teams.

III.2 Metrics and measurement

1. Because we have many features, we need to implement dimension reduction using clustering(substitute some feature with one single level) , PCA, or just add regular term like using LASSO. Then we can evaluate the effect of this step by checking the percentage of variance being explained or general performance of our model on validation set.
2. Because our data is not in a large scale, we can use the 10 folds cross validation to run our model 10 times to make it more convincible.
3. When implementing classification, we can use the confusion matrix to analyze the wrong classified cases.
4. We can use graphs like AUC and ROC to make judgement of our model's performance, and adjust the threshold when using perceptron.
5. Because of the model's property, we can find the probability for prediction in logistic classification. The SVM can also generate the prediction probability according to the distance to hyperplane, which is useful to our evaluation.
6. When we use clustering, we can use within-group deviation as a metric.

We can use graph to make above statement less abstract:(That's not in our training set, just to show some data pattern)



The pair plot of WorldCup features with different labels

IV Resource

We use public dataset from *http://www.football-data.co.uk/data.php*

And we just use standard computer device, python and scikit-learn document to implement our project.

V Contributions

- Richard Yang is responsible for establishing model , part of essay and program writing.
- Zhoujingpeng Wei is responsible for graph decoration, model evaluation and finding reference.
- Jiaxin Hu is responsible for data searching, data preprocessing, essay correction and explaining the visualization .