

# 1.Introduction to the problem

As one of the most famous sports, Football is played by approximately 250 million players in over 200 countries and has billions of fans worldwide. With the vigorous development of football games, every club has its own fans. This makes predicting game results one of the topics that football fans pay the most attention to. There are three possible outcomes for a game: home win, draw and away win. Due to its popularity and the small number of possible outcomes of the game, predicting outcomes is a very interesting and seemingly simple challenge. However, it is difficult to predict the final result because the way the team plays on a specific date depends on many factors, such as the current format, the last team meeting, confrontation, offensive and defensive skills, the personal abilities of key players and even the psychology influence of fans in the stands.[1]

Machine Learning, as a multi-disciplinary interdisciplinary, involving probability theory, statistics, approximation theory, convex analysis, algorithm complexity theory and other disciplines. The purpose of machine learning is to build a special algorithm (rather than a specific algorithm) that allows the computer to learn from the data to make predictions.

In this project, we aim to use various algorithms within machine learning to build a model for predicting the football games results and find the better one from the comparison.

## 2.Literature Review

Football is a game with few goals, which means that many inconspicuous factors can determine the final result. Therefore, it is quite difficult to choose features and classification methods that help predict. In the paper[2], authors list the steps of the system design to build the prediction model.

### 2.1 Feature Selection

Many factors such as the form of the football team, home advantage, the overall quality of the team, and the personal qualities of the players in the team will affect the final result of the football game. The problem of selecting features can be solved in two ways [3]. The first way implies that we have a certain understanding of the problem, and based on this knowledge, we choose those characteristics that we think will most affect the final result. The other way means that we have little or no knowledge of the problem, so we select all the features that may affect the results, and then gradually try to determine the most influential features[4].

## 2.2 Selecting Learning Algorithms

In order to achieve better prediction results, various algorithms are considered to determine which provides better results.

### Bayesian networks

Bayesian networks, BNs, provide a means for capturing, displaying, and making available in a usable form knowledge, often obtained from experts in a given field[5].

When approaching a new problem there are two techniques which are commonly used. The first assumes we have some idea how the situation under investigation works, construct a model, and using this model select the attributes believed to contribute to the effect under investigation. An example of this approach to this type of problem is given. The second approach assumes little knowledge of the underlying mechanisms involved so we look at all the probably relevant attributes and try to determine those which have the most significant effect. This is still in effect the construction of an a priori model, but only a very informal one[6].

### LogitBoost

The LogitBoost algorithm is a boosting algorithm. The Boosting algorithm consists of a set of weak classifiers. If they are used as independent classifiers, their performance will be poor, but the combination will work well.

### K - nearest neighbor algorithm

It is the representative algorithm from the group of lazy classifiers. Classification of new examples is done by finding k nearest neighbors in the space of features (typically using Euclidean distance) from the examples in the learning set. Based on these examples voting or another method determines classification of new examples[7].

## 2.3 Conclusion

Predicting the results of football matches poses an interesting challenge due to the fact that the sport is so popular and widespread. However, predicting the outcomes is also a difficult problem because of the number of factors which must be taken into account that cannot be quantitatively valued or modeled. As part of this work, a software solution has been developed in order to try and solve this problem. During the development of the system, a number of tests have been

carried out in order to determine the optimal combination of features and classifiers. There is room for further improvement, primarily in the area of feature selection. If we were to model the form for each and every player in the match we could probably achieve better results. This way we could monitor each player's form during the season and determine its influence on the final score. Besides, Larger data set for learning would also help to predict future outcomes[8].

## 7. Other Ideas

### Picture recognition

With the development of search engines, people are more reliant on the help of the program to find out the content they need, not only words, but also pictures. We can distinguish the cat pictures and dog pictures, but computers recognise things differently. We need to teach them with lots of pictures so that they can figure out the difference. But when the algorithm learns enough data and well programmes, it can do things that humans can not, such as the texture or position of the picture.

### Language Translation

With the development of globalization, there is more communication within people from different countries. The huge difference between cultures leads to the problem which is the improper translation. People can normally recognise sentences translated by translators because of the different structures and choice of words. We think that it can be improved with the addition of filters or tags. From our perspective, if we can add a tag system to stuff we want to translate, then train the algorithm with more exact context, then it should come out with more authentic results.

### Virtual assistant

Siri, Alexa, Cortana, guess we are familiar with at least one of them. Setting alarm clock, counting time, making a call, virtual assistant follow our command by recognizing our voice and certain words. We are interested to find out the way of how voice can be recognized.

## 8. Discussion of Topics

## Machine Learning

As the chosen method of the project, we have explained a lot about it. It is the core of artificial intelligence and the fundamental way to make computers intelligent. The research directions of traditional machine learning mainly include research on decision trees, random forests, artificial neural networks, and Bayesian learning.

### Decision Trees

Decision trees are based on the known probability of occurrence of various situations, by constructing a decision tree to obtain the probability that the expected value of the net present value is greater than or equal to zero, evaluate project risk, and judge its feasibility. A decision tree is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category. In our project, it can be used for linking between individual factors, such as the strength of the offense of a player can lead to a chance of the goal.

### Markov Decision Processes

MDP is used to simulate the achievable randomness strategy and rewards of the agent in an environment where the system state has Markov properties. MDP is a mathematical model that simulates the random policy and rewards of an agent in the environment, and the state of the environment has Markov properties. It contains 5 model elements, state, action, policy, reward and return. In our project, it can be used to simulate the strategy of a team in different situations. But due to the complexity of the reinforcement learning we may not use this method in the project.

### Game Theory

Strictly speaking, game theory is mainly a mathematical model of conflict and cooperation between rational decision makers. From my perspective, it is a study of how to play various games in this world reasonably. Traditional machine learning is mostly regarded as an optimization problem. What we need to do is to find an algorithm that can search for the optimal solution. At this time, the idea of game theory is very important. Generalization is sometimes referred to as "structural risk minimization."

