CSE 4108

Artificial Intelligence Lab

Winner Runners up Prediction Classification

Project Report

Submitted by

Md. Siam Ansary ID: 14.02.04.104

Lab Group: B2

Group no. : 06

Ahsanullah University of Science and Technology
Department of Computer Science and Engineering
Fall 2017

Project

Winner Runners up Prediction Classication

Description

Football is a very popular worldwide sport. Our project resolves around the different football leagues around the world. It is a classifying problem of predicting the Winners or Runners Ups.

We have applied six classifiers on the dataset. Unimportant columns were not brought into consideration for classification. As the feature columns were of different domains, they were feature scaled. Through cross validation, training was done and later testing was executed. As it is a binary classification problem, dummy values were created to handle the codes better.

Dataset

The dataset of this problem has been created with the information from the site **World Football** (http://www.worldfootball.com). There are 120 entries in the dataset with fourteen columns. We use the records of four popular League records. The columns of the dataset are as such in Table 1.

Table 1: Columns of the dataset

Columns	Description
YearFrom	The year a season starts
YearTo	The year a season ends
Club	Name of the football club
Country	The country the club is from
League	The league of participation
Pld	Number of games played in the particular season
W	Number of games won in the particular season
D	Number of games drawn in the particular season
L	Number of games lost in the particular season
GF	Goals scored by the club
GA	Goals scored against the club
GD	Goal difference
Points	Points gained by the club in the particular season
Outcome	If the club was winner or runner up that season

The target column is **Outcome**.

Used classification models

Six classifier models were used. Two of them are from ensemble models. The used models are:

- **Logistic Regression**
  Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The hypothesis is given as below:

  $$h_\theta(x) = \frac{1}{1 + e^{(-\theta)^T \times x}}$$

  In a binary classification problem, if 0 and 1 are two classes, for hypothesis to be less than 0.5, it predicts 0 and else predicts 1.

- **Stochastic Gradient Descent Classification**
  Stochastic gradient descent works well with a large dataset. It does not need to look at all training data in one iteration, rather a single training data in one iteration. The technique is efficient

Table 2: Source modules of different classifiers

Source module	Classifier
sklearn.linear_model	LogisticRegression
sklearn.linear_model	SGDClassifier
sklearn.neighbors	KNeighborsClassifier
sklearn.tree	DecisionTreeClassifier
sklearn.ensemble	AdaBoostClassifier
sklearn.ensemble	RandomForestClassifier

and easy to implement.We use SGD Classifier that implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties for classification

- **K Nearest Neighbors Classification**
  The k-nearest neighbors algorithm is a non-parametric method used for classification and regression.The input consists of the k closest training examples in the feature space.
  In classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

- **Decision Tree Classification**
  The decision tree classification technique is organized a series of test questions and conditions in a tree structure. The root and internal nodes contain attribute test conditions to separate records that have different characteristics. All the terminal node is assigned a class label.

- **Adaptive Boosting Classification**
  Adaptive Boosting can be used in conjunction to improve performance. The output of the other learning algorithms is combined into a weighted sum that represents the final output of the boosted classifier.It begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

- **Random Forest Classification**
  Random forests operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set

Adaptive Boosting and Random Tree classification techniques are of ensemble learning method. Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

Comparison of performance

To show the comparison of models used, we use accuracy score, precision score, recall score and f1 score.

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to the all observations in actual class. F1 Score is the weighted average of Precision and Recall.

As in the code, cross validation is used, at different run, different entries are included in training set and

Table 3: Comparison of models at a particular run

Classification model	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.766666666667	0.923076923077	0.666666666667	0.774193548387
SGD	0.733333333333	0.692307692308	1.0	0.818181818182
K Neighbors	0.566666666667	0.692307692308	0.5	0.58064516129
Decision Tree	0.633333333333	0.769230769231	0.555555555556	0.645161290323
Ada Boost	0.666666666667	0.75	0.666666666667	0.705882352941
Random Forest	0.7	0.846153846154	0.61111111111	0.709677419355

test set. Also different models have different advantages and disadvantages in different data distributions. So, a model may work well in one run but may work bad in a different run.

In table 3 we include the accuracy , precision , recall and f1 scores of the models used of a run to indicate comparison in their performance on the dataset.

Discussion

The problem is of binary classification and two neighbouring entries are of opposite classes. Logistic regression works best in this type of problem. SGD classification may give good result in some runs but also may perform absolutely worst as it looks at a single training data in one iteration. Decision tree and random forest classification can give moderate predictions. Adaptive boosting and K nearest neighbors do not give good results. K neighbors on average works worst as the neighbors are equally distributed of opposite classes. On a whole, for this problem and dataset,Logistic Regession technique is best suited.