Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017)
IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9

Team Strategizing using a Machine Learning Approach.

Vignesh Rao
Department of Computer Science and Engineering
Visvesvaraya National Institute of Technology
Nagpur, India
Email: raovignesh0210@gmail.com

Aman Shrivastava
Department of Mechanical Engineering
Indian Institue of Technology
Roorkee,India
Email: aman.srivastava999@gmail.com

***Abstract*—Team strategizing is an important aspect which requires critical analysis to ensure a desirable near-optimum performance. The key to solve this issue is by tapping the available talent within the team which at times, can be elusive. With increasing competition, a talented team, with an ineffective and outdated scouting strategy, may have to face unfavourable results. In this paper, we have conducted research in the domain of Sports, specifically Soccer. Strategy considered in the research is centered around deciding the lineup of a team by assessing the skillset of the players. Considering the novelty of the approach, we have developed our own web scraping algorithm to collect the dataset. Machine Learning models like Neural Network(MultiLayer Perceptron), Random Forests and Logistic Regression have been used to make predict the position a particular player will perform best at. The accuracy of the said models have been analysed for comparative analysis.**

***Keywords*—Neural Network, Random Forests, Logistic Regression, Machine Learning, Team Strategizing.**

I. Introduction

There have been several instances in history where the introduction of analytics and statistics completely revolutionized the field of sports. One of the most notable being Sabermetrics, introduced by Billy Beane, as the manager of Oakland Athletics. The data driven approach employed by Mr. Beane and his team, led to Oakland A's 20 game unbeaten record during the 2002 MLB season. In our research, we have used statistics and advanced analytics to aid in the scouting process for football teams and deciding the playing lineup based on individual players performance statistics. Some professional statistical analysis firms like Prozone Sports[6] and Opta Sports[7] exist nowadays which provide data to the football clubs, coaches and leagues.

For building a truly strong team, it is imperative that players playing at every position will give their best performance at that position. This classification, traditionally done manually, can be done using statistical models beings used today to make the task faster and more optimized. For the analysis, we trained machine learning models using the data developed by Electronic Arts for the latest edition of their FIFA game franchise.

It has been established that the use of data from the FIFA franchise has several merits that traditional datasets based on historical data do not offer. Since 1995 the FIFA Soccer games provide an extensive and coherent scout of players worldwide. For each attribute, we have an integer from 0 to 100 that measures how good a player is at that attribute. Examples of attributes are: dribbling, aggression, vision, marking and ball control. Observe that it seems to be unfeasible to accurately characterize players in these attributes automatically. Thus, all of those are gathered and curated by the company whose job is to bring the gameplay closer to reality as possible, hence preserving coherence and representativeness across the dataset.

The FIFA 18 dataset that has been used for this analysis provides statistics of about 17000 players on about 70 different attributes, 30 of which, we propose, are relevant for position classification. These attributes are optimal indicators to determine the performance of a player at a particular playing position. We used state-of-the-art machine learning classification algorithms to classify players at the Attacking, Mid and Defence positions.

Post extraction of the data and forming a clean dataset, Machine Learning models like Neural Network, Logistic Regression and Random Forests have been used. The train data with labels is fed into the classifier. After the training phase, input to the trained classifier is the test data whose output class need to be predicted. Various performance metrics like F1 score, precision, recall and accuracy have been calculated. Confusion Matrix has been plotted for a better insight of the performance of the model.

The construction of the remaining part of the research paper is as follows. A concise description of the dataset and the Machine Learning models are presented in Section II. Section III deals with the overall Methodology of the research paper. Experimental results with a short discussion is presented in Section IV. A condensed Conclusion is provided in Section V.

II. Background

A. Dataset

A scraping algorithm is developed to form our dataset by extracting data from a website named: https://sofifa.com. The player personal data alongwith ids and most importantly their playing and style statistics were collected for further analysis. The dataset comprises of around 70 features of which only 30 relevant features were filtered out. The total count of players in the dataset is around 17,800. Some key features include aggression, agility, balance, ball control, sprint speed, long passing, finishing, standing tackle,interceptions, etc. 80% of

the retrieved data is utilized for the training and the remaining 20% is for the testing phase.

B. Artificial Neural Network

An artificial neural network is the collection of individually connected components called artificial neurons. Processing and transmission of signals takes place between the neurons. The weights associated with the neurons varies with learning which can either increase or decrease. The ANN used in our research is MLP(MultiLayer Perceptron). The basic structure of MLP is shown in Fig 1.

1) MultiLayer Perceptron: It is the class of feedforward Artificial Neural Network. The minimum number of layers in this case is three. The technique utilized for training is back propagation which is a supervised learning.[1] Non linearly separable data can also be distinguished by MLP.

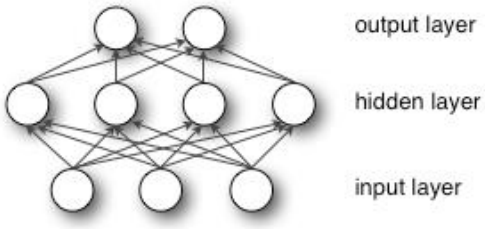

Fig. 1: MultiLevel Perceptron

The basic case consisting of a single layer hidden layer is expressed mathematically as follows:

$$g(x) = b + W \tanh(c + Vx)$$

Where, x is the input vector,
V is a matrix of input to hidden weights.
W is the matrix of hidden to output weights.
c is the bias vector.

The computation involving the 'tanh' function is the output of the hidden layer. The operation when performed again on top of it gives rise to an addition of another layer. Stacking of such layers results in building up of deep neural network.

C. Logistic Regression

Logistic Regression is a regression model where the dependent variable is categorical. The dependent variable can be binary or multi-class. Thus there are mainly two types based on the type of the dependent variable.

1) Binary Logistic Regression

2) Multinomial Logistic Regression.

In our case the dependent variable is multiclass and thus multinomial logistic regression is used. The input is the set of independent variables which might be binary,categorical,etc.

D. Random Forests

Random Forests are an ensemble learning method for tasks like classification,regression,etc. It does so by building multiple decision trees at the training phase. In case of classification, the final output of the classifier is the mode of the classes predicted by the individual decision trees. Whereas in case of regression, mean is taken into consideration. Due to its ability to deliver on large sets of data, it is scalable. It is generally used in outlier detection and replacing the missing data. It is a supervised classification and is introduced well in [2]

III. Methodology

The ultimate goal of our approach is to assign optimal position to the players depending on their skillset. In this case, three output classes are pre-decided: attack, mid and defense. The entire flow of the process can be observed in Fig. 2. The approach taken in this is quite modular in its organization.

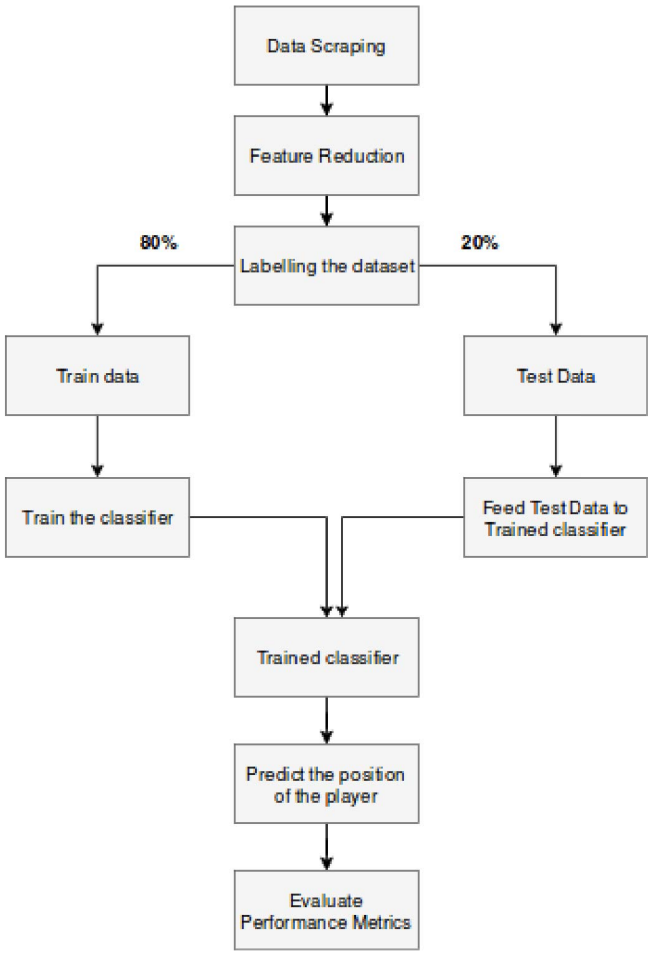

Fig. 2: Complete Flow of the procedure

The overall flow of the process can be expressed as follows:

1) Data Scraping:
The process starts with data scraping from the website which has statistics of around 17000 players with approximately 70 features. The attributes include the players playing style

statistics and personal information. The dataset collected is stored in a csv file for further processing.

2) Feature Reduction:
The next step involves feature reduction. Since there are many features that are not relevant to deduce our results, we can drop them. Thus, the selection of 30 relevant features is done for improving the accuracy of the model by supplying quality data to the classifier. For example, attributes like personal information are futile for training the classifier and thus can be ignored for analysis.

3) Labelling the dataset:
The dataset has a column where the preferred position of the player is stated. A total of 14 positions are then mapped to the 3 predecided classes. Data normalization is performed on all the features of the dataset except the attribute having the preferred position to ensure consistency. Thus the value of each feature lies between 0 to 1.

After cleaning the dataset, 80% of the data is randomly allocated to train the classifier and the remaining 20% is used for testing.

4) Training the classifier
The machine learning models used in this approach are Neural Network(MultiLayer Perceptron), Random Forests and Logistic Regression. Using GridSearchCV, optimal performing neural network is selected based on the value of alpha and number of hidden layers. The optimal number of hidden layer and the value of alpha comes out to be 20 and 0.001 respectively. These parameters are used in training the Neural Network. In case of Logistic Regression , Multinomial Logistic Regression is used since the dependent classes is a multi-class. Random forests is used with default parameters. The trained classifier is stored in a pickle file. The pickle file is used to store the classifier object in a serialized way for further use in the testing phase. This ensures that training is done only once and is decoupled from testing.

5) Testing phase
After training the model, test data is fed in and the trained classifier is loaded from the pickle file. The desired output class is determined,in this case-the optimal position of the player. The output of the testing phase is then provided for the analysis.

6) Evaluating Performance Metrics:
The output from the classifier is evaluated based on some performance metrics like F1 score, Precision, Recall and accuracy. The primary metric is the F1 score. For visualization, a confusion matrix is also plotted as shown in Fig 3,4 and 5.

IV. Results

The performance metrics for a classifier include F1 score, accuracy, precision and recall with F1 score being our primary measure, and rest secondary. F1 score considers both precision and recall and it is the harmonic mean of these two measures. Mathematically, these metrics can be represented as below:

Precision: Is the ratio of relevant instances to the retrieved instances[3].

$$P = \frac{TP}{TP + FP} \tag{1}$$

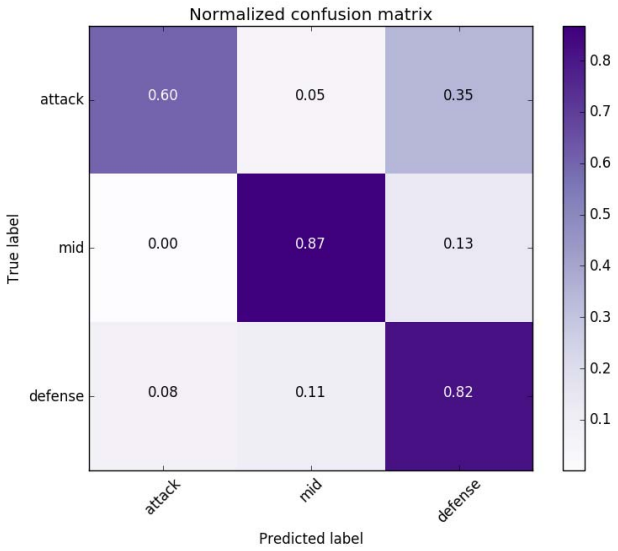

Fig. 3: Confusion Matrix for Neural Network

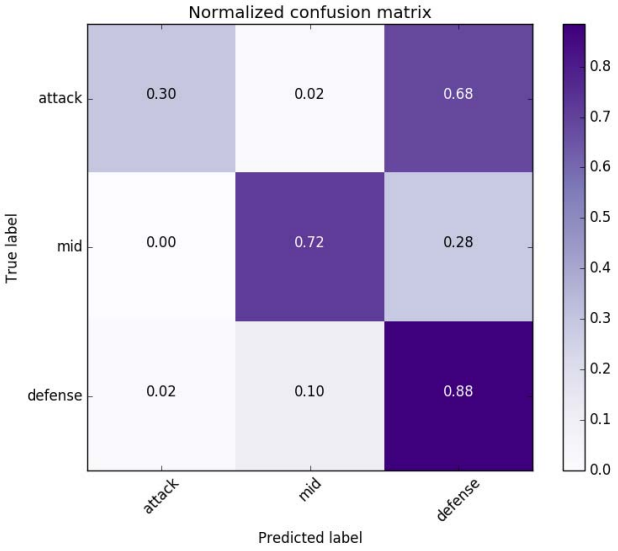

Fig. 4: Confusion Matrix for Logistic Regression

where, TP stands for True Positive and FP denotes False Positive.

Recall: Is the fraction of relevant instances retrieved to the total number of actual relevant instances[3].

$$R = \frac{TP}{TP + FN} \tag{2}$$

where, TP stands for True Positive and FN stands for False Negative.

F1 score: Is the harmonic mean of precision and recall[4][5]. Since it considers both precision and recall into one value, it is better to use this as a primary metric.[8]

$$F1 = 2 * (\frac{P * R}{P + R}) \tag{3}$$

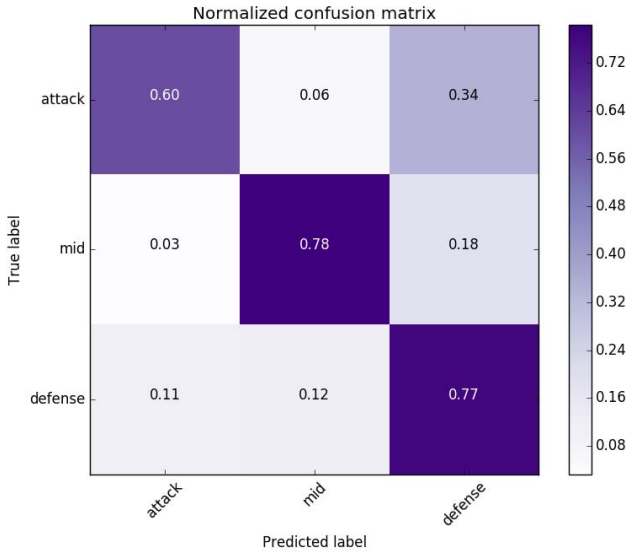

Fig. 5: Confusion Matrix for Random Forests

where, P is the Precision and R stands for Recall.

Evaluating the results, it can be pointed out from Table 1 that Neural Network(MultiLayer Perceptron) has performed the best with an accuracy of 79.01% and F1 score of 0.787. Random Forests has performed well too with F1 score of 0.739 and an accuracy of 74.07%. Logistic Regression with an accuracy and F1 score as 71.92% and 0.697 respectively lies at the bottom of the table with respect to the metrics taken under consideration. On close inspection of Confusion matrix it can be observed that all the three classifiers have correctly predicted the output for the mid and defense positions to a great extent. Whereas, the accuracy with which it has predicted the attack position is quite less compared to other positions. The less accuracy in the attack position is due to the fact that in most of the formations, less players play in that position in comparison to mid and defense. The accuracy of the all the three models are not that significant owing to the fact that players might have qualities that can suit them to play in more than one position. There is further scope in this research to improve on the accuracy and also to scale it on other domains.

TABLE I: Results

Performance Metrics				
Model	Precision	Recall	F1 score	Accuracy
Neural Network	0.788	0.790	0.787	79.01%
Logistic Regression	0.747	0.719	0.697	71.92%
Random Forests	0.739	0.74	0.739	74.07%

V. Conclusion

In this research, machine learning techniques have been harnessed to achieve an effective team strategizing analysis. Neural Network, Random Forests and Logistic Regression are the models used in our paper. The results delivered cannot grant us a human level accuracy in predicting the position of the player depending on his skillset. Examination of the performance metrics reveal that Neural Network has performed the best amongst the other models. Further research on this untouched area might lead to an increased accuracy. In Future, this approach can be scaled to other domains like Education,Business,etc for tapping the available talent and making the optimal use of it.

References

[1] S.N Sivanandam, S.Sumathi, S.N.Deepa, *"Introduction to Fuzzy Logic Using MATLAB"*, Springer Berlin Heidelberg, New York.

[2] Leo Breiman, *Random forests*, Machine Learning. vol. 45, no. 1, pp. 532, 2001.

[3] T. Landgrebe, P. Paclk, R. Duin, and A. Bradley, *Precision-recall operating characteristic (P-ROC) curves in imprecise environments*, in Proceedings of ICPR, 2006.

[4] J. Davis and M. Goadrich, *The relationship between precision-recall and ROC curves*, in Proceedings of the 23rd International Conference on Machine Learning, ser. ICML 06. New York, NY, USA: ACM, 2006, pp. 233240 [Online]. Available: http://doi.acm.org/10.1145/1143844.1143874

[5] C. Rijsbergen, *Information Retrieval*,2nd ed. London: Butterworths, 1979.

[6] Prozone sports. [Online]. Available: www.prozonesports.com

[7] Opta sports. [Online]. Available: www.optasports.com

[8] Y. Baeza and B. R. Neto, *Modern Information Retrieval.* Boston, 1999