Machine Learning in Table Tennis Match Prediction

Sophie Chiang
The Perse School
Rouse Project

***Abstract*—**

- **Using historical data to predict the outcome of a table tennis match**
- **12 features are extracted and derived to capture the quality of players and their opponents**

I. Introduction

- Motivation, opener, What is the point in predicting the outcome of a match?
- Despite the popularity of the sport, little work has been done in prediction
- Machine Learning algorithms can be used to identify patterns and relationships in player statistics to evaluate which are important to determine the outcome of a match
- Some methods that have been used in tennis prediction has been applied
- Limitations, limited number of data

II. Related Work

- Machine Learning, supervise learning, classification
- Knottenbelt [5] proposed a common opponent model by analysing match statistics for opponents that both players have encountered in the past, providing a fair basis of comparison, and can compute the probability of each player winning a point on their serve, and hence the match.
- Voeikov et al. [8] present a neural network TTNet that is able to extract temporal and spatial data from processing high resolution table tennis videos, potentially capable of substituting manual data collection by sport scouts, and can assist in referee decision making.
- Number of works in tennis prediction [4], [7].

III. Dataset

- Table tennis match data was retrieved from a data set available on OSAI https://osai.ai/
- This includes match results and statistics from the Tokyo 2020 Olympics, as well as Tischtennis-Bundesliga, the top professional German table tennis league
- heatmaps of ball percentage in different regions of table

IV. Feature Engineering & Selection

- Forehand, backhand, long rally and short rally winning percentages
- New features were derived by taking the differences in percentages between a player and their opponent, such as serve advantage or the difference in rank ($D_i = R_i - R_j, D_j = R_j - R_i$) [7]
- This was done to achieve symmetry and to avoid any inherent *bias*
- Two rows were used to represent a match, one from the perspective of each player
- Normalisation and standardisation of data
- Player statistics were combined to derive new ones [1]; for example *balance* is the average of serve, long rally and forehand advantage, and represents a player's overall skill compared to their opponent
- A *Long rally* is defined to be a rally with over five shots

V. Method

- Binary classification problem; a win and a loss is represented by 1 and -1 respectively in perspective to a player
- Multiple machine learning algorithms used
- Logistic Regression [6], Random Forest Classification [2], Multi-Perceptron Neural Network Layer Models, Support Vector Machines
- 10-fold cross validation was used to evaluate the *accuracy* of a model
- The data set is split into 10 folds; 9 of these was used as a training set and one as a testing set
- Each sample is used the same number of times for training and only once for testing and the *accuracy* results is calculated by taking the average of every validation cycle

VI. Experimental Results

- Accuracy can be compared across different models, as well as with and without newly derived features (ablation study)
- Learning curve graphs for logistic regression and SVM
- 'Feature importance' from a random forest model can be extracted, computed as the mean decrease in node impurity for that feature across all trees [3]

VII. Discussion

- Confusion matrices on true positive and true negatives etc
- Compare results across different models and evaluate each
- Hyper-parameter tuning for each model
- Train, test and validation set

VIII. Conclusion

- Further Work

References

[1] T. Barnett and S. R. Clarke. Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, 16(2):113–120, 2005. 1

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 1

[3] A. P. Cassidy and F. A. Deviney. Calculating feature importance in data streams with concept drift using online random forest. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 23–28. IEEE, 2014. 1

[4] A. Cornman, G. Spellman, and D. Wright. Machine learning for professional tennis match prediction and betting. *Stanford Unverisity*, 2017. 1

[5] W. J. Knottenbelt, D. Spanias, and A. M. Madurska. A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers & Mathematics with Applications*, 64(12):3820–3827, 2012. 1

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 1

[7] M. Sipko and W. Knottenbelt. Machine learning for the prediction of professional tennis matches. *MEng computing-final year project, Imperial College London*, 2015. 1

[8] R. Voeikov, N. Falaleev, and R. Baikulov. Ttnet: Real-time temporal and spatial video analysis of table tennis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 884–885, 2020. 1