A Review of Data Mining in Sports

Rahul Raghavendra Huilgol
Indian Institute of Technology Guwahati
Guwahati, India
h.rahul@iitg.ernet.in

Simrat Singh Chhabra
Indian Institute of Technology Guwahati
Guwahati, India
simrat@iitg.ernet.in

ABSTRACT
The vast amount of data that the field of sports provides has only recently been tapped into by data mining researchers. This paper looks at popular data mining techniques and how they have been used for various purposes in the area of sports. The applications of Artificial neural networks, Decision trees and Fuzzy systems are discussed in detail.

General Terms
Sports, Data mining, Artificial neural networks, Decision trees, Fuzzy system

1. INTRODUCTION
Fans and experts alike have since long indulged in predicting and arguing with each other about the outcomes of sports matches. With the large amount of data available (especially since the advent of Internet), it was only natural that statisticians and computer scientists will take interest in finding out patterns and making prediction using that data. Coaches and team managers of various sports could benefit from such analysis and use it to make their strategies. Sports betting is another domain in which such techniques could make a significant improvement in the accuracy of their odds.

The field of sports has huge amounts of data in the form of game videos, audio and text commentary and statistics of players and teams. Most of the data has been collected in recent years as technology has advanced. This data presents a huge potential for data mining techniques to extract patterns. But this is not without its challenges. This data is also unstructured and noisy. The data can have lot of features which may not be relevant to the task. Also, unlike some other fields there is no standard dataset which can be used to compare the performances of different techniques.

In this paper, we review the various data mining techniques which have been used for result prediction and event detection in sports. We have also analysed the accuracy of each technique and what applications they have been used in. There are also other techniques such as Naive Bayes Classifier, Support Vector Machines and Logistic regression which have been used to extract relevant information in the field of sports [10]. However, these methods are not as widespread and haven't been very successful and hence we haven't covered them in this review.

2. NEURAL NETWORKS
Artificial neural networks are computational models inspired from the natural nerve cells in our brains. They are adaptive systems which associate weights for each of the connections in the network. The output of the network depends on these weights which are trained by minimizing the error between the actual output and the output of the network. Neural networks are capable of automatically learning the hidden dependencies of historical data by learning the weights and using this to predict the future. This coupled with the fact that they have been shown to be robust to noise have made them popular models for prediction in sports. There are many types of neural networks differing in their architectures. Feed forward neural networks only allow flow of information in the forward direction. Recurrent neural networks have loops which allow information to flow back. Researchers have tried various models with the aim to model sports data.

Multi layer perceptrons are feed forward neural networks which consists of multiple layers of neurons. McCabe, Trevathan, 2008 [9] used a multi-layer perceptron to predict the outcome of games given some basic information. They used data from several major leagues sports like the Australian Rugby League, National Rugby League and the English Premier League. A set of features were designed like Home and away performance, performance in previous game, performance in previous n games, Points-for, Points-against, Overall performance, team ranking, etc. A three layer MLP with one unit for each input feature was used. The output of the network was normalized to get a value between 0 and 1, where a value close 1 indicated that the model predicts that the team will win. The team with the higher output value in a particular game was taken as the predicted winner. The proportion of predictions which matched the actual result was taken as success rate. Predictions were made for each round of the season and the model was retrained after

each round using all examples till that particular point in time. Human experts typically only have a success rate between 60% and 65% at this task. Their system named the "McCabe's Artificially Intelligent Tipper" (MAIT) achieved high accuracy with the best case accuracies being 68% in AFL and 75% in Super Rugby. There were hardly any other systems doing similar analysis to compare against so the authors decided to compare MAIT's performance against human tipsters. MAIT took the first position in the international tipping competition TopTipper in the 2006-2007 season. MAIT had an accuracy of 93.8% in the Rugby World Cup of 2003 and 83.3% accuracy in 2007.

Z. Ivankovic et al. (2010) [8], also used feedforward neural networks to analyze basketball games. They used data from the Serbian Basketball league. Features selected included the percentage of succesfull one point throws from the free throw line, 2 point and 3 point throws from six different positions on court, defensive rebound, offensive rebound, steal, etc. After training on 75% of the data, it was evaluated on 25% of the data. Apart from showing predictions of upto 80% accuracy, they could also point out which feature had the most influence on the result of the game. They used the weights learnt by the network for a particular feature as a measure of how much influence that feature has on the result.

Davoodi, E. , Khanteymoori, A.R. (2010) [7] also showed that neural networks perform well in the context of horse race prediction. They used features like type of race, number of horses in the race, track condition, horse weight, horse's jockey, horse's trainer, race distance and weather data from horse-races in the Aqueduct Race track in New York, USA. The symbolic data variables were encoded into continuous ones. They trained one neural network for each horse, where each neural network outputs the predicted time for the horse to finish the horse. The architecture was chosen using the method of network growing. Here, starting from a small network neurons are added gradually increasing the size of the network. Out of all those architectures, the one which minimized the mean square error of the networks was chosen, giving a network with 4 layers and each neuron is connected to every neuron in the next layer. They analyzed and compared the performance of five training algorithms to train the network, namely Backpropagation, Backpropagation-with-momentum, Conjugate Gradient Descent and Quasi-Newton, Levenberg-Marquardt. All the training algorithms were shown to produce similar accuracy of around 77%, with the backpropagation algorithm giving slightly better results although it was slower than the rest. The Levenberg-Marquardt algorithm was shown to be the fastest.

Wickramaratna, K., Chen, M., Chen, S.C. and Shyu, M.L. (2005) [15] introduced a special neural network based framework to detect goal events in videos of soccer. The event of goal is a rare event, number of goal shots is less than 1% in the dataset. For rare events, a single neural network performs badly because the backpropagation algorithm converges very slowly when most of the examples belong to one particular class [1]. The authors introduce a simple approach to overcome this by employing an ensemble of neural networks each of which is trained using bootstrapped sampling and the predictions of each network are combined to give the final prediction. Bootstrapped sampling is used to overcome the imbalanced dataset by creating sets of training samples which have comparable amount of samples from both the classes. Training ensembles of multilayer perceptrons allows us to use all of the training data and reduces generalization error. Sollich, P. and Krogh, A. (1995) [13] showed that when training with ensemble neural networks overfitting is actually useful. So this ensemble is under-regularized so as to overfit thus giving a 100% accuracy for each of these component networks. Because of overfitting each of these networks will not perform as well on the rest of the data. The output of the whole system is proposed to be the weighted combination of all the networks. The component network with the least generalization error is assigned higher weight when combing the predictions. The bootstrap sampling ratios of goal to non-goal events were experimented with 1:1, 1:2 and 1:3. The ensemble with bootstrapped sampling was shown to be better than a single neural network in all cases. The recall value increased significantly with only a small decrease in precision. As the goal to non goal ratio decreased, there was a decrease in recall and increase in precision. The results showed that this was an effective way to detect goal in soccer videos.

3. FUZZY SYSTEM

While traditional logic operates with binary values (True / False), fuzzy logic, as its name suggests, allows for variables to take intermediate values. So instead of a simple 0 or 1, a fuzzy variable might take values in between. Fuzzy logic (and fuzzy algorithms eventually) came about as a way to model human thinking. Humans hardly ever think in strict binary values, but instead tend to prefer a gradual transition from one end to the other. Fuzzy algorithms are sequences of instructions in which the variables and the conditional statements can both take fuzzy values.

Fuzzy algorithms have been used in many data mining applications such as intrusion detection [5], web usage mining [3] and pattern recognition [4] .

Trawinski [14] used various fuzzy algorithms for result prediction in basketball. WEKA (Waikato Environment for Knowledge Analysis) was used for feature selection. The author considered features such as whether it was a home game or an away game, the previous results of both the teams, and the number of points they had scored in their earlier games. 15 such features were initially taken. Eight different algorithms such as GainRatioAttributeEval and ConsistencySubsetEval were then selected for getting the features for each algorithm. On these selected features, he applied 10 different fuzzy algorithms using KEEL (Knowledge Extraction based on Evolutionary Learning) and compared the results obtained by each. LinearLMS had the beset accuracy of 66.7%. To improve upon the result, the author modified the features considered and added features such as average of scored points were game, average of gained assists per game, and average valuation per game. This increased the accuracy of the now best algorithm Clas-Fuzzy-Chi-RW to 71.5%.

Rotshtein, Posner and Rakityanskaya [12] applied fuzzy methods to predict results of football matches. They constructed

a fuzzy model considering five varibale : number of matches which classify as big win, small win, draw, small loss and big loss. Fuzzy knowledge matrices were constructed with common-sense reasoning. To apply the knowledge matrices, generalized fuzzy approximator was used. To tune the prediction model by changing the values for the weights for the fuzzy rules, the authors used a genetic algorithm and a neural network. The system got an accuracy varying from 83.3% to 94.6%.

4. DECISION TREES

Using decision trees for predictive modelling has been used extensively in data mining [2] [11]. Decision trees have been found useful as their logic is visible and can be explained even to a layman. In decision trees, splits are made based on particular features and the majority class value is computed at the leaves. Splits for which those values are high (when run on training data) are then used for testing.

Chen, Shyu, Chen and Zhang [6] used decision trees for soccer goal detection. Their work was divided into 3 major parts: video parsing, data pre-filtering, and data mining. They used video features like pixel_change (average percent of changed pixels between frames within a shot), histo_change (histogram difference between frames) and others. They also used audio feature like volume and energy. They used 10 audio features and 5 visual features to represent each segment of video and audio. Since the number of goal shots are a very small percentage of the total shots, the authors applied data-filtering based on observation and prior knowledge. This helped in reducing 81% of video shots.

For the data mining part, the authors used C4.5 decision tree algorithm. They applied 2-way split at each level and used information gain ratio criterion to determining the appropriate attribute for partitioning. The chosen method achieved a recall of 90.2% and precision of 94.9%.

5. CONCLUSION

Sports mining is a recent field and there is still huge scope for improvement. The current state of most sports involves a high degree of professionalism, and in this competitive environment, even a slight extra insight into the variables which go into deciding a match can give a team or a player that competitive edge over their rivals. Players and teams can plan their strategies using analytics provided by data mining techniques.

Improvement in event detection methods can lead to easier classification and archiving of video and audio data of matches. Such methods could also be used for generating automatic highlights.

Apart from direct applications in sports mining, new methods applied in data mining in sports can later be generalised to other applications as well.

6. REFERENCES

[1] M. C. R. S. Anand R1, Mehrotra KG. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks and Learning Systems*, 1993.

[2] C. Apté and S. Weiss. Data mining with decision trees and decision rules. *Future generation computer systems*, 13(2):197–210, 1997.

[3] S. Araya, M. Silva, and R. Weber. A methodology for web usage mining and its application to target group identification. *Fuzzy sets and systems*, 148(1):139–152, 2004.

[4] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981.

[5] S. M. Bridges and R. B. Vaughn. Fuzzy data mining and genetic algorithms applied to intrusion detection. In *Proceedings of 12th Annual Canadian Information Technology Security Symposium*, pages 109–122, 2000.

[6] S.-C. Chen, M.-L. Shyu, M. Chen, and C. Zhang. A decision tree-based multimodal data mining framework for soccer goal detection. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 1, pages 265–268. IEEE, 2004.

[7] K. A. Davoodi, E. Horse racing prediction using artificial neural networks. *Recent Advances In Neural Networks, Fuzzy Systems and Evolutionary Computing*, 2010.

[8] M. M. B. R. D. Ivankovic, Z. Rackovic and M. Ivkovic. Analysis of basketball games using neural networks. 2010.

[9] A. McCabe and Trevathan. Artificial intelligence in sports prediction. 2008.

[10] D. Miljkovic, L. Gajic, A. Kovacevic, and Z. Konjovic. The use of data mining for basketball matches outcomes prediction. In *Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on*, pages 309–312. IEEE, 2010.

[11] L. Rokach. *Data mining with decision trees: theory and applications*, volume 69. World scientific, 2008.

[12] A. P. Rotshtein, M. Posner, and A. Rakityanskaya. Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, 41(4):619–630, 2005.

[13] P. Sollich and A. Krogh. Learning with ensembles: How overfitting can be useful. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 190–196. MIT Press, 1996.

[14] K. Trawinski. A fuzzy classification system for prediction of the results of the basketball games. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pages 1–7. IEEE, 2010.

[15] C. M. C. S. Wickramaratna, K. and M. Shyu. Neural network based framework for goal event detection in soccer videos. 2005.