Artificial Intelligence in Sports Prediction

Alan McCabe
MAIT Technologies
Belfast, Ireland
Email: alan@mymait.com

Jarrod Trevathan
School of Mathematics, Physics and Information Technology
James Cook University
Email: jarrod.trevathan@jcu.edu.au

***Abstract*— This paper presents an extension of earlier work in the use of artificial intelligence for prediction of sporting outcomes. An expanded model is described, as well as a broadening of the area of application of the original work. The model used is a form of multi-layer perceptron and it is presented with a number of features which attempt to capture the quality of various sporting teams. The system performs well and compares favourably with human tipsters in several environments. A study of less rigid "World Cup" formats appears, along with extensive live testing results in a major international tipping competition.**

I. Introduction

This paper extends previous work [8] in exploring the utility of neural networks, in particular multi-layer perceptrons, in predicting the outcome of sporting contests given only basic information. This is a less traditional application area of neural networks and somewhat of a novelty, however, the basic principles of machine learning still apply. Additionally, attaching to predictions an indication of how certain the predictor is, and rewarding such predictions appropriately, are important issues in many fields.

The data used in this work was taken from several different sources and covers four major league sports: the Australian National Rugby League (NRL), the Australian Football League (AFL), Super Rugby (Super 12 and Super 14) and English Premier League football (EPL) from as early as 2002. Each of these leagues has different characteristics, schedules, lengths and team structures (for example EPL has a feature in which the bottom three teams are "relegated" at the end of each season and replaced by three new teams).

The data contains noise in that there are obviously details influencing the contest outside of those which are being captured in the feature set. Firstly, there is what is often referred to as individual "form" of the players, however it is believed that in most cases, the team's overall skill level will transcend poor individual form. Secondly, there is the fact that the team's skill level, or quality, can be affected by the unavailability of "star" players due to injury, suspension or representative duties (such as in the NRL where players may be called away from their team for state or country duties). Some experiments were performed on trying to account for player availability and these are reported later in the paper. There was a conscious effort to ensure that there was no subjectivity in the feature set.

The paper is organised as follows: Section II gives a background on the neural network engine used to make the predictions; Section III describes the raw data used and the feature extraction process; Section IV details the experiments conducted and the results of the work, including comparisons to "expert" tipsters; future work is presented in Section V and Section VI contains the concluding remarks.

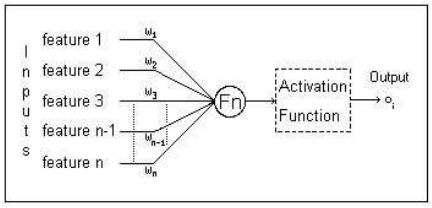

Fig. 1. A simple neural network topology, with features, weights and a single output unit.

II. Modelling the Feature Space

A. Neural Networks

The main reason for using neural networks (NNs) to model the feature space is the model's ability to learn the relationship between inputs and outputs upon presentation of examples [3, 13]. It is necessary only to provide a set of sample data (also known as training data or a training set) to the network and the use of learning (or training) algorithms such as back-propagation perform an adjustment of the network to better model the problem domain.

There are many types of neural networks and one of the most popular models is the multi-layer perceptron (MLP). MLPs associate a weight with each of the input features (see Section III-A for a discussion of the features used in this study) according to that features importance in the model (see Figure 1). This type of network topology is imminently suited to the well-defined domain discussed in this paper, because several features exist and a weighting must be associated with each according to its contribution to the solution. These weights can be set to specific initial values (possibly to facilitate an intentional bias) or simply randomly assigned. The learning algorithm then adjusts the weights to minimize the error between the target output (the desired output provided in the learning examples) and the actual output (the output as calculated by the MLP). There are a number of learning algorithms that can be used to optimize the weights in MLPs, such as back-propagation, conjugate gradient descent and Levenberg-Marquardt [3]. The two learning algorithms used in this work were back-propagation and the conjugate-gradient method, both classical algorithms which are effective, relatively simple and well understood. Some argue that other

978-0-7695-3099-4/08 $25.00 © 2008 IEEE
DOI 10.1109/ITNG.2008.203

learning algorithms often perform faster [4], but as this study deals with a small feature set with no requirement for real-time operation, the above-mentioned advantages outweigh any perceived increase in speed.

The back-propagation and conjugate-gradient methods work by iteratively training the network using the presented training data. On each iteration (or epoch), the entire training set is presented to the network, one case at a time. In order to update the weights, a cost function is determined and its derivative (or gradient) with respect to each weight is estimated [12]. Weights are updated following the direction of steepest descent of the cost function. A common cost function and the one used in this work is the root mean squared (or RMS) error. During experimentation with the two methods, back-propagation was a little slower to learn than the conjugate-gradient approach, but both methods resulted in almost identical error rates, with back-propagation slightly more accurate. Further discussions of MLPs and the two learning algorithms can be found in most neural network texts, for example [3, 12, 13].

III. Input Data

The raw data for this work's experiments was gathered from several sources, and depending on the league being examined there were between thirteen and thirty-eight rounds of competition. From this raw data it is possible to determine round by round statistics (features) for each team including their current success rate, recent performance, the points they've scored in the competition to date (to rate their offensive capabilities), the points scored against them (to rate their defensive capabilities) and several other indicative features.

A. Feature Extraction

A conscious effort was made to exclude any subjective features from the raw data. Features obtained were based solely on details such as scoreline, recent performance and position on the "league ladder" relative to other teams. This removes the need for any human judgement in generation of features and removes any bias (intentional or otherwise) on the part of that human. A set of features was obtained for each team, for a given round of competition, as follows:

Points-for: the total points scored by the team in matches so far this season.

Points-against: the total points scored against the team in matches so far this season, expressed as a negative number.

Overall Performance: the team's performance based on their win/loss record. Two points are awarded for a win (or three points in the case of EPL), one point for a draw and no points for a loss. Performance is then the sum of these values over each round of competition so far.

Home Performance and Away Performance: the cumulative performance value calculated using only home games and only away games respectively. The Overall Performance feature can hide specific details such as home-ground performance, for example if the team has a 90% success rate at their home ground and a 10% success rate away from home, then an overall success rate of 50% hides some important information when trying to predict a winner.

Performance in Previous Game: the team's performance in their most recent game. In the first round this value is typically taken from the previous season's final game.

Performance in Previous n Games: the average of the performance for the most recent n games. Up to five previous games were considered in the feature set. This is an attempt to gauge the recent form and take into account whether the team is on a winning or losing streak.

Team Ranking: the team's position on the league ladder based on a list of the teams, sorted by their overall Performance value. This feature's use is obvious as, with all other things being equal, a team with a high ranking is expected to defeat a team with a low ranking.

Points-for in Previous n Games: the average of the points scored by the team in the most recent n games. Values for n of one to five were used as five separate features.

Points-against in Previous n Games: the average of the points scored against the team (expressed as a negative number) in the most recent n games. Values for n of one to five were used in the feature set.

Location: a value indicating whether the current game is played at the team's home venue or elsewhere. The value 1 is taken for a home game, and 0 for an away game.

Player Availability: not in the basic feature set, this feature was used in further testing reported below. A "star" player is defined as one who is currently involved in their nation's national side (and/or state side in the NRL competition). This feature then is the proportion of star players who are unavailable in a given week.

When appropriate, the feature values were averaged over the number of matches played. This was done so that a meaningful comparison may still be made between two teams which have played a different number of matches.

IV. Experimentation

The experiments conducted involved extracting the aforementioned features, and then constructing the models. As mentioned in Section II the choice was made to use a multi-layer perceptron to model the features, and back-propagation (which proved slightly more effective than the conjugate-gradient method) to facilitate learning. Specifically, a three-layer MLP was used with nineteen input units, or twenty when Player Availability is included (one for each feature), ten hidden units and a single output unit. The output unit was normalised to be a value between zero and one inclusive.

The feature set values were calculated for each team for each round of competition. The MLP was trained using all examples from previous rounds, going back to the previous season if necessary, and re-trained after each round. Predictions were made for the current round by using the MLP to calculate an output value for each team based on that team's feature set. An output value of close to one for a particular team indicated a high level of confidence that the team was going to win their upcoming match, and an output value closer to zero indicated a lower confidence level.

The output values for the two teams competing in each game were calculated and the team which had the highest output

value (i.e., the highest confidence that the team would be victorious) was taken as the predicted winner (or tip) for that match. Success rates were then calculated as the proportion of tips for which the predicted winner matched the actual winner.

A. Results

The results presented here are an augmentation and update of those presented in the original article [8]. Averages and highlights for the four different league competitions are presented, as well as comparisons to human tipsters and the result of two World Cup format tests. In addition, the results of brief experiments on player availability adjustments are presented. All predictions were made available under the name of "McCabe's Artificially Intelligent Tipper" or MAIT on the official system website [7] several days in advance of the matches. In addition, the predictions were often announced in print media, radio, or other websites [1, 15, 16].

It should be noted that it is often difficult to measure the success of a system such as this, as there are few benchmarks for the use of neural networks in this domain. Systems with a similar structure have been used to perform other predictive tasks, such as use of neural networks to predict prices on the stock market [6, 9, 10, 11, 17]. These systems do not typically perform with outstanding success and are rarely more effective than a naive human investor (largely due to the fact that stock market prices are affected by a very large number of variables, many of which are difficult to quantify).

A small number of other published systems exist which attempt to apply neural networks or other logical methods to the sporting arena, with varying levels of success [2, 14]. The best-case result for a single season reported in purely NN systems was a 58% success, which occurred in the Australian National Rugby League. Regular (human) experts typically have success rates of somewhere between 60% and 65% for NRL, AFL and Super Rugby. In the EPL, human experts successfully predict somewhere between 50% and 55% of matches, where the much higher prevalence of draws results in lower overall accuracy. Note that for the purposes of results presented here, unless a draw was specifically predicted, an actual result of a draw is counted as incorrect.

B. League Competitions

All four of the league competitions mentioned previously have been subject to live testing for at least the last three seasons. The NRL predictions for example have been made publicly available since 2002. Table I presents the best and worst whole-season performances, along with the average accuracy for each of the target sports. It should be noted that that the NRL accuracy was well down in 2007 (resulting in the worst performance recorded, 11% lower than the previous worst), where further inspection has identified the fact that the system had not been applying any home ground advantage, which will require continued analysis and monitoring in future.

The Super Rugby competition was slightly different in that two new teams were introduced in 2006, which necessitated some small changes in the algorithm. The weights for the NN models for the two new teams were set at an average of all

TABLE I
LIVE TESTING ACCURACY FOR THE FOUR TARGET SPORTING LEAGUES.

League	Best	Worst	Average
AFL	68.1%	58.9%	65.1%
NRL	67.2%	52.2%	63.2%
Super Rugby	75.4%	58.0%	67.5%
EPL	58.9%	51.8%	54.6%

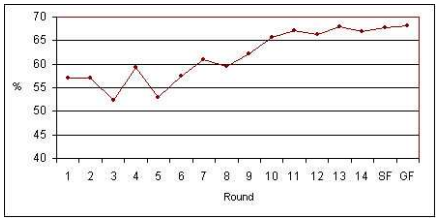

Fig. 2. Performance over the 2007 Super 14 season. After an early period of adjustment, the accuracy steadily and consistently improved.

weights for the models for other teams. It was surprising how quickly the models learned the new teams details, normalising performance within two to three weeks. As can be seen in Figure 2, the typical behaviour of Super Rugby aligns with what is classically expected from artificial intelligence algorithms, with an initial performance roughly equivalent to random assignment of results, steadily improving throughout the remainder of the season. The average performance in Super Rugby has remained quite high at 67.5%.

English Premier League football (also known as soccer) posed a new problem with such a high prevalence of draws (23% of EPL matches in the last three years have ended in a draw). As a result, an allowance must be made to specifically predict draws, which is done by specifying a tolerance level for the difference between ratings for the two teams. If the difference falls below this tolerance level, a draw is predicted, otherwise the team with the higher rating is the predicted winner for that contest. Over the last two seasons, 35 draws have been predicted, with 16 of those being correct at a 45.7% success rate, approximately double what would be expected with random assignment.

C. Comparison to Human Tipsters

As mentioned, it's difficult for an effective comparison given the absence of a large number of other systems doing similar analysis. A more meaningful test of the prediction algorithms is to compare the performance against a large set of human tipsters. To this end, the MAIT system was entered in a major international tipping competition called TopTipper [5] in the 2006-2007 season. This competition hosts thousands of (human) contestants from year to year, of varying skill levels. Figure 3 illustrates the MAIT system's performance versus the human competitors, showing how the relative performance steadily improved throughout the season. By the final week of the competition the system had taken first position, which was a considerable achievement.

D. World Cup Format

The extension to "World Cup" environments presented a significant challenge to the system. The major difference between World Cups and more structured league formats is

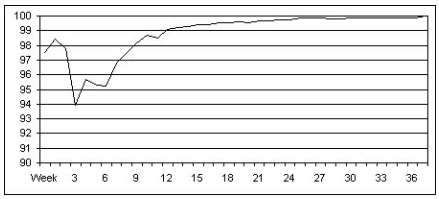

Fig. 3. Performance in TopTipper's English Premier League tipping competition. This figure tracks the percentile that the MAIT system lay in (the percentage of other competitors that the MAIT system was ahead of).

that each of the teams have a very different performance history on which to draw. All teams have played a different number of games at greatly different intervals and against greatly different opposition.

In these cases it was necessary to considerably expand the number of teams under consideration, so as to have a measure of the "quality" of all teams involved in previous matches, in order to assess the significance of a given team's previous results. At the beginning of the tournament proper, the ratings for all non-active teams were disregarded and the process continued as with a normal league format.

This procedure was followed for both the 2003 and 2007 Rugby World Cup tournaments, where live predictions were again made. In 2003 the Sydney Morning Herald newspaper presented the predictions in direct competition with their own resident human expert. The final result for the MAIT system was 45 correct predictions from the 48 matches (at a 93.8% success rate), compared to 42 from the human expert. In the 2007 tournament, one in which several more unusual results presented themselves, the system still performed quite well, recording 40 correct predictions from the 48 matches (83.3%).

E. Player Availability

During the 2006 NRL season, experiments were performed with a simple player availability algorithm. A requirement was imposed that all features be entirely objective. Also included was the definition of a "star" player as one who was involved in the most recent national or state side for which they were eligible. The number of star players unavailable for a given team for a given round was then used as an input feature in the model (divided by seventeen, the total number of players nominated for a single team), and the models re-trained using data from the 2004 and 2005 seasons. A total of nine differences (compared with models not incorporating the player availability feature) occurred over the course of the 2006 season, with a single extra correct prediction being made when player availability is taken into account.

Despite the single extra hit, an improvement of just over 0.5%, this approach was abandoned in 2007 for two reasons. Firstly, significant effort was required to maintain the list of star players, as well as performing comparisons with nominated teams on a weakly basis. Secondly, the actual prediction often changed (sometimes repeatedly) in the days leading up to the game based on the injury status of these star players.

V. Future Work

There are several possibilities for future directions with this work, the most imminent being an extension into further sporting arenas. The major professional American sports seem an obvious choice, and will allow a comparison with a larger set of existing systems. Expansion to different sports also allows for the development of a richer feature set, and the monitoring of the effects of these on the models.

The other focus area for future work in the short term is margin prediction. That is, not only are the models going to be used to predict a winner, but also the margin of victory, which obviously poses a more significant challenge to the system.

VI. Conclusions

Despite an existing "novelty" value for this work, there is still theoretical interest in the modelling of features in a noisy environment and the use of machine learning techniques to predict probabilistic events. Perhaps the primary attraction of sport in general is that there are so many elements which contribute to the result, and that on any given day either team is capable of winning. This same fact is what makes the prediction process so difficult and why so much time and money is spent by individuals trying to predict winners in various tipping competitions and gambling outlets.

This paper described an extension to previous work in the generalization and modelling of behaviour of teams in sporting contests. Results were reported for different sports and various seasons, and compared against human "expert" tipsters. The multi-layer perceptrons used were able to adapt very quickly and perform well despite the limited information and the outside influences not included in the feature set.

References

[1] *ABC Radio National*. http://www.abc.net.au/rn/, 2007.
[2] Baulch, M. *Using Machine Learning to Predict the Results of Sporting Matches*. Thesis, University of Queensland, 2001.
[3] Bishop, C. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
[4] Hassoun, M. *Fundamentals of Artificial Neural Networks*. Massachusetts Institute of Technology Press, 1995.
[5] Internet Digital Media Australia. *TopTipper: Online Tipping Competitions*. http://www.toptipper.com/, 2007.
[6] Kalyvas, E. *Using Neural Networks and Genetic Algorithms to Predict Stock Market Returns*. Masters Thesis, University of Manchester, 2001.
[7] McCabe, A. *McCabe's Artificially Intelligent Tipper (MAIT)*. http://www.mymait.com/, 2007.
[8] McCabe, A. *An Artificially Intelligent Sports Tipper*. Artificial Intelligence '02, Canberra, 2002.
[9] McCann, P. J. and Kalman, B. L. *A Neural Network Model for the Gold Market*. http://citeseer.nj.nec.com/308853.html.
[10] McNelis, P. D. *Neural Networks in Finance: Gaining Predictive Edge in the Market*. Academic Press Advanced Finance Series, 2004.
[11] Op 't Landt, F. W. *Stock Price Prediction Using Neural Networks*. Maters Thesis, Leiden University, 1997.
[12] Pessoa, L. *Multilayer Perceptrons versus Hidden Markov Models: Comparisons and Applications to Image Analysis and Visual Pattern Recognition*. Qualifying Examination Report, Georgia Institute of Technology, 1995.
[13] Russell, S. and Norvig, P. *Artificial Intelligence - A Modern Approach (2nd Edition)*. Prentice Hall, 2002.
[14] *Swinburne Sports Statistics*. http://www.swinburne.edu.au/sport/, 2007.
[15] *Sydney Morning Herald - Business News, World News and Breaking News*. http://www.smh.com.au/, 2007.
[16] *Townsville Bulletin: Local and Regional News*. http://www.townsvillebulletin.com.au/, 2007.
[17] Toulson, D. L. and Toulson, S. P. *Use of Neural Network Ensembles for Portfolio Selection and Risk Management*. NeuroCOLT Technical Report Series, NC-TR-96-046, 1996.