ISSN 1750-9823 (print)
International Journal of Sports Science and Engineering
Vol. 04 (2010) No. 03, pp. 131-138

Data Mining in Sport: A Neural Network Approach

John McCullagh

Faculty of Education, La Trobe University, Bendigo, Australia

(*Received December 23, 2009, accepted March 21, 2010*)

Abstract. Data Mining techniques have been applied successfully in many scientific, industrial and business domains. The area of professional sport is well known for the vast amounts of data collected for each player, training session, team, game and season, however the effective use of this data continues to be limited. Many sporting organisations have begun to realise that there is a wealth of untapped knowledge contained in their data and there is an increasing interest in techniques to utilize the data. The aim of this study is to investigate the potential of neural networks (NNs) to assist in the data mining process for the talent identification problem. Neural networks use a supervised learning approach, learning from training examples, adjusting weights to reduce the error between the correct result and the result produced by the network. They endeavour to determine a general relationship between the inputs and outputs provided. Once trained, neural networks can be used to predict outputs based on input data alone. The neural network approach will be applied to the selection of players in the annual Australian Football League (AFL) National Draft. Results from this study suggest that neural networks have the potential to assist recruiting managers in the talent identification process.

Keywords: Data Mining, Talent Identification, Neural Networks.

1. Introduction

Data Mining involves using mathematical or statistical tools and techniques for extracting knowledge from large amounts of data. Numerous studies have demonstrated successful outcomes using data mining techniques to estimate various parameters in a variety of domains [1-4]. While there have been many success stories, the application of data mining techniques in the sporting domain has been limited. The following studies describe some of the data mining applications conducted in the sporting domain. The applications take a variety of forms including evaluation of game strategies, prediction of training loads, injuries, team and individual performance, as well as talent identification in various sports.

Data mining techniques were used on freshmen cadets at the United States Military Academy for analysing physical aptitude data as a predictor of future physical performance [5]. The programs produced several similar association rules and identified performance standards that incoming cadets should achieve in order to accomplish the strenuous physical requirements of the service academy. The results demonstrated the potential for data mining techniques to predict future performance from large volumes of historical data. Tschopp et al. [6] conducted a study to evaluate the predictive value of physiological, medical, psychological, anthropometric, social and personal characteristics for medium term success in junior elite soccer players. Height, isokinetic strength of the knee flexors and age at entry to the football club were statistically significant predictors. The study also concluded that the implementation of a multidisciplinary assessment of elite junior players would be most effective at age 15 due to greater homogeneity with increasing age.

Neural networks have been used successfully in college football as an alternative to the techniques used to rank football teams [7]. The neural network model offered enough flexibility in its parameter settings to take into account many of the different factors involved in the decision making process. Advanced Scout software was used to seek out interesting patterns in basketball game data. The information derived proved successful in assisting coaches to assess game strategy as well as formulate game plans for future games [8].

Pyne et al. [9] investigated the relationships between anthropometric and fitness tests from the Australian Football League (AFL) Draft camp and the career progression of these players in AFL Football. The results demonstrated that the 20 m sprint, jump, agility and shuttle run tests are a factor with the career progression of AFL footballers. The US National Football League Combine is the NFL's equivalent to the AFL National Draft. Players are tested on a variety of measures and the results are presented to all clubs to assist in the

drafting process. McGee et al. [10] conducted a study to assess the relationships between the player test results and draft status and demonstrated that the Combine results can be used to predict accurately the draft status of certain positions (running backs, wide receivers and defensive backs). Other positions were not accurately predicted. The research findings could be used to determine which position an athlete is most suited to and in which position an athlete is most likely to be successful. The study also found that the first and second round drafted athletes were taller, heavier and faster over 10, 20 and 40 yards as well as scored higher in agility runs, vertical and broad jumps.

The aim of this study is to investigate the applicability of neural networks to predict the future playing ability of players in the AFL National Draft from all forms of data available including anthropometric, psychological and skill assessment.

2. Neural Networks

Neural networks are one of the data mining techniques used when large amounts of data are available. Neural networks are mathematical models which can be used to model complex relationships between inputs and outputs or to find patterns in the data. A backpropagation neural network will be used in this study.

2.1. Backpropagation Neural Networks

A backpropagation neural network is a multi-layered non-linear feed-forward network trained by the backpropagation learning algorithm. It contains an input layer, one or more hidden layers and an output layer. All nodes (other than the input nodes) generate an internal activation which is calculated by summing all of the input weight products as outlined in Equation 1.

$$\text{Activation} = \sum_{j=1}^{n} W_i X_i + W_B \tag{1}$$

Where: $X_1 .. X_n$ are the inputs

$W_1 .. W_n$ are the associated connection weights

W_B is the bias weight.

A bias acts exactly as a weight on a connection from a node whose activation is always one. Increasing the bias increases the net input to the unit [11]. The structure of the network including the number of nodes used, their organisation into layers and the connections between them is referred to as its architecture (see Figure 1). The most common architecture is the fully interconnected multi-layered network. The backpropagation neural network is trained by being presented with numerous examples where each example consists of a set of inputs and their corresponding output(s). The "learning" that takes place in backpropagation neural networks occurs by the adjustment of the connection weights [12].

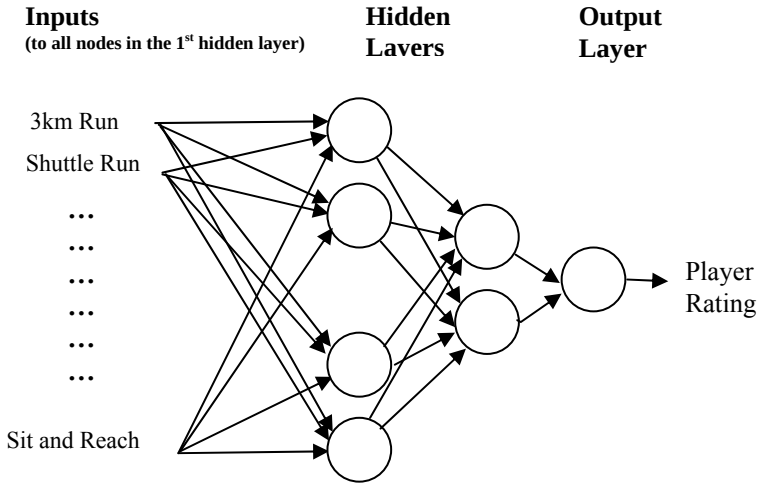

Figure 1: Neural Network showing the inputs and output for the AFL draft problem

Many factors influence the success that can be achieved by a backpropagation neural network on a particular problem including the architecture and the parameters. The architecture chosen for a network is influenced by the complexity of the mapping between inputs and outputs. A complex mapping will often require a greater number of hidden nodes in the architecture. A number of parameters must be specified before the backpropagation neural network can be used, including the momentum, learning rate, initial weight size, epoch size and the number of passes. The weights in a backpropagation neural network indicate the strength of the connections between neurons and represent the learning that has been achieved [13]. The number of passes indicates how many times the training data is presented to the network. The network may be set to run the full number of passes or an error criterion may be used to stop the network.

The neural networks used in this study as individual classifiers or as part of a modular neural network will be backpropagation neural networks.

2.2. Modular Neural Networks

The modular neural network (MNN) approach is to decompose the problem into simpler sub-tasks, each one being handled by a separate module. The modules are then combined to produce an overall solution, demonstrating a very natural way to solve a complex problem [14,15]. A neural network ensemble is a type of modular neural network which consists of a set of whole problem classifiers whose individual decisions are combined in some way to classify new examples. The success of the ensemble relies on the individual classifiers producing accurate classifications, as well as being distinctly different in their error patterns. A committee of these networks does not improve performance if all errors of the individual networks coincide. On the other hand, if all errors of individual networks do not coincide and the performance of the networks is maintained, the committee can demonstrate a considerable improvement in performance. Therefore, it is beneficial to make the errors among the networks in the committee less correlated in order to improve the committee performance. Neural networks with less correlated errors are called diverse networks [16,17]. Numerous methods exist to develop the individual networks, however, they all have the desired aim to create a number of high performing networks where there exists a low correlation between the errors for each network [14].

Neural network ensembles, applied to the AFL draft problem, will be developed in this study.

3. Experimental Design

3.1. Player Data and Ratings

Table 1. Player ratings and descriptions

Rating	Description
≥8	Elite AFL player
7	Very good AFL player
6	Good AFL player
5	Plays a majority of games in the seniors (≥ 80%) and is regarded in the top 22 in the team
4	Just outside of the top 22 in the team. Plays < 80% of games in the seniors
3	Plays a majority of games in the reserves. Not thought of as a regular senior AFL player at this stage.
2	Unlikely to become a regular AFL player. Minimal or no AFL games
1	Drafted but no impact
0	Not drafted

The Australian Football League (AFL) conducts an annual draft camp where promising U18 footballers are tested over a variety of physiological, medical, psychological, skill, anthropometric, social and personal characteristics. The results from this camp are used to assist recruiting managers to make player selections in the AFL National Draft which occurs in November each year. The player data set used in this study consists of 386 player examples and 58 input attributes. The player attributes include body composition, flexibility, anaerobic and aerobic power, visual tests, TAIS (Test of Attentional and Interpersonal style), psycho-motor tests, skill assessments and subjective assessments on strengths, weaknesses and personal attributes. The 58 inputs were selected as they were the attributes tested across all years between 1999 to 2005. Skill

assessments such as marking, kicking and clean hands were assessed numerically by coaches. Other attributes such as big game ability, coachability and aggressiveness were quantified by the use of a set of criteria to analyse the coaches written reports for each player [18]. Players were given a rating out of 10 to indicate their value in the AFL competition. The ratings used in this study were the average of three subjective assessments from people experienced in the AFL drafting process. Table 1 was used as a guide to assist in this assessment. Players were required to be in the AFL system for at least 3 years for an assessment to be made of their ability.

A second set of ratings were used for the classification experiments. The classes are outlined in Table 2.

Table 2. Classes used for experimentation.

Class	Description	Rating
GOOD	Range from: plays the majority of games in the seniors to elite	≥ 5
AVERAGE	Range from: not a regular senior player to no impact at all	< 5

3.2. Neural Network Parameters

The neural networks were trained using the parameters specified in Table 3 below. The parameters were derived by conducting a series of trials involving varying parameters and assessing the effect on the NN's output. Once trained the networks performance was assessed on the testing set to provide a measure of performance on unseen data.

Table 3. NN parameters

Parameter	**Value**
Architecture	58 - 5 - 1
Learning rate	0.3
Momentum	0.2
Epochs	2000

3.3. Neural Network Ensembles

Neural network ensembles were used to trial their effect on overall performance. The problem space was sub-divided into two categories as outlined in Table 4 and an ensemble was developed for each category.

Table 4. Categories used for neural network ensembles.

Height	**Category**
>=188 cm	Key Position Player (KPP)
<188 cm	Mid Sized Player (MSP)

One ensemble of five neural networks was developed to specialize in players for the KPP category, while a second ensemble was developed to specialize in players for the MSP category. Specialist neural networks were developed using targeted training sets. KPP specialist networks were trained on 80% examples from the KPP category while MSP specialists were trained on 80% examples from the MSP category. The five networks in each set were each trained using a different random seed to enable some form of diversity to be achieved across the networks. A simple majority voting system was used to achieve an overall classification.

4. Results and Discussion

Three sets of experiments were conducted in this study. The first set of experiments, called the Regression Experiments, used a numeric rating between 1 and 10 as the NN's output. The second and third sets of experiments, referred to as the Classification Experiments, used two rating classes, GOOD and AVERAGE, as the output. The second set of experiments used neural networks while the third set of experiments used neural network ensembles to make their classifications.

In order to compare the results for the NNs to the performance for the recruiting managers (RMs), the following assumption was made: the actual draft order for each year is based on an accumulation of

knowledge from recruiting managers across the 16 AFL clubs, even though each club works independently. The RMs' predictions used in this study is the draft order for each year.

4.1. Regression Experiments

Table 5 outlines the performance of the recruiting managers and the neural networks on the testing data. The correlations are calculated as an average across the two years of the testing data (2004 & 2005).

Table 5. Correlations for RM versus rating and NN versus rating for the testing set.

	Correlation
RM v Rating	-0.41
NN v Rating	-0.27

The result presented for the recruiting managers (RM v Rating) is the correlation between the actual draft order for a particular year and the current rating for a player. It would be expected that a higher draft order (i.e.: 1, 2, 3 etc) would result in a higher rating and vice versa, resulting in a negative correlation. The results presented for the NN (NN v Rating) is the correlation between the NN's predicted draft order and the current rating for each player. The NN's predicted draft order is calculated from the NN player rating output for each player. Both the RMs and NNs correlations were negative as expected, however the recruiting managers clearly outperformed the NNs' predictions.

4.2. Neural Network Classification Experiments

Table 2 outlines the two player classes used in experimentation. To allow a comparison to be made between the NNs' classification performance and the recruiting managers, the following reasoning has been used. An analysis of the draft data indicates that on average approximately 20 players each year achieved a rating of 5 or above. This information was used as the basis to develop an RM classification for each player in the draft. As previously stated, the draft order each year is taken as an estimation of the accumulation of knowledge from the recruiting managers across the 16 clubs. Therefore if each player selected in the top 20 is assigned a GOOD rating, and players selected below 20 an AVERAGE rating, this may be used as an estimation of the RMs' classification [18].

Table 6 shows the performance of the recruiting managers and the NN across the entire testing set, while Table 7 shows the respective percentage correct for GOOD predictions.

Table 6. RM and NN predictions on the testing set.

Scenario	**Percentage Correct**
RM predictions	66.8%
NN predictions	60.1%

Table 7. Performance of RMs and NNs GOOD predictions on the testing set.

Scenario	**Percentage Correct**
RM predicts GOOD	44.1%
NN predicts GOOD	35.9%

The performance of the recruiting managers clearly outperforms the performance of the neural networks. The RMs' predictions are 11.1% higher than the NN on the entire testing set, and 22.8% higher for GOOD predictions. This would be expected when we take into account the inputs that are used by each system. While the neural networks rely only on the draft camp data, the recruiting managers have extensive networks of people viewing games of potential players over many years, interviews, videos etc as well as the draft camp data.

Further analysis of the GOOD predictions shows that while there was a certain amount of disagreement between the RMs' and the NNs' classification, there were also numerous times when they agreed. Table 8 shows the percentage correct for GOOD predictions where the recruiting managers predicted correctly, the

NN predicted correctly and both predicted correctly.

Table 8. GOOD predictions on the testing set: 3 scenarios.

Scenario	Percentage Correct
RM predicts GOOD, NN disagrees	29.5%
NN predicts GOOD, RM disagrees	17.1%
RM and NN predict GOOD	49.8%

The results show that the RM clearly outperforms the NN on the percentage correct of GOOD predictions. In the cases where the RM and the NN both agree on their GOOD prediction, the percentage correct was 49.8%. This compares favourably with the results presented in Table 7 where the RM predictions are 44.1% correct and the NN 35.9% correct for GOOD predictions. The results demonstrate a potential for the neural network to be used as an additional source of information to assist in the decision making process.

4.3. Neural Network Ensemble Classification Experiments

The neural network ensemble technique uses the power of multiple neural networks to sub-divide a problem in order to assist in improving classification accuracy. As key position players commonly demonstrate a different set of characteristics to mid-sized players, it was determined that a modular approach to handle each group of players would be well suited to this problem.

Figure 2 demonstrates the process used to derive a player rating estimation for each set of player data. The rating is a classification as outlined in Table 2. The height criteria (Table 4) is used to sub-divide the problem into two sections. This results in a player rating classification being made by either the KPP or MSP ensemble. Majority voting is used to obtain a consensus opinion from the five neural networks in each ensemble.

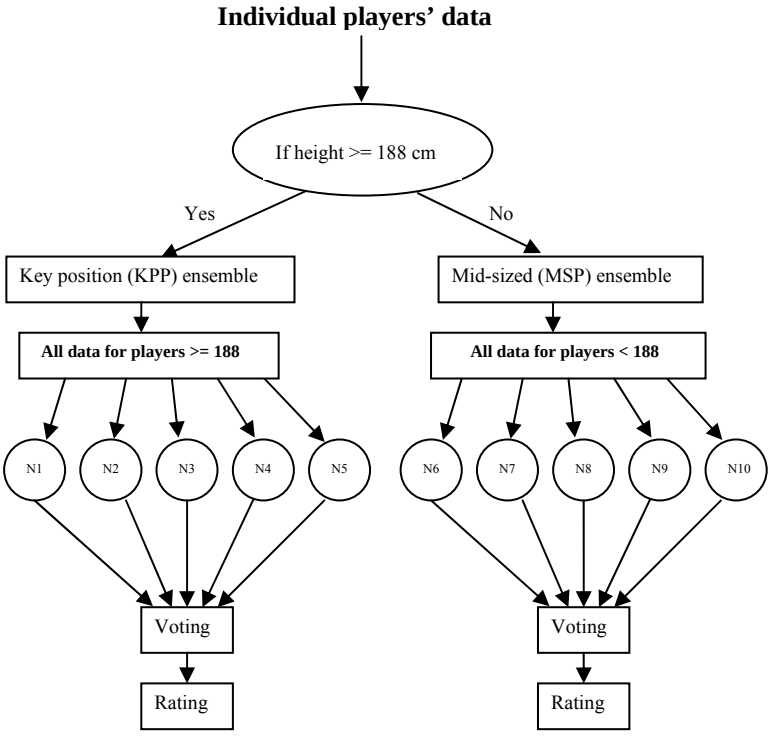

Figure 2: Process to estimate player ratings from neural network ensembles.

Table 9 compares the results for the recruiting managers and the NN ensembles across the entire testing set, while Table 10 shows the respective percentage correct for GOOD predictions.

Table 9. RM and NN ensemble predictions on the testing set.

Scenario	**Percentage Correct**
RM predictions	66.8%
NN ensemble prediction	61.3%

Table 10. Performance of RM and NN ensemble GOOD predictions on the testing set.

Scenario	**Percentage Correct**
RM predicts GOOD	44.1%
NN ensemble predicts GOOD	37.5%

The results demonstrate that the recruiting managers outperformed the neural network ensembles. However, the performance of the neural network ensembles showed an improvement over the performance of the single neural network. The ensemble estimations demonstrated a 2.0% and 4.4% improvement over the single neural network's performance on the entire testing set and the GOOD predictions respectively.

Table 11. GOOD predictions on the testing set: 3 scenarios.

Scenario	**Percentage Correct**
RM predicts GOOD, NN ensemble disagrees	29.2%
NN ensemble predicts GOOD, RM disagrees	18.5%
RM and NN ensemble predict GOOD	52.3%

The results in Table 11 demonstrate the potential for using neural network technology to assist in the draft selection process. A percentage correct of 52.3% was achieved when the recruiting manager and the neural network ensemble agreed with GOOD player predictions. This compares favourably with a percentage correct of 44.1% for the recruiting manager's GOOD predictions.

While the low numbers used in these experiments make it difficult to draw conclusions, it is interesting to note that on their own NNs do not approach the success achieved by the recruiting managers, however, they may have potential to be used as a secondary source of information to confirm or improve a decision.

5. Conclusion

A vast amount of data is available to recruiting managers to assist in the draft selection process. However, techniques to efficiently use this data continue to be an area of concern. The success of recruiting managers' decisions in the National Draft can have a major impact on the on-field and off-field success of a club in future years. As a result, it is essential that the clubs are able to make the best selections possible.

The results from this research indicate that the recruiting managers clearly outperform the neural networks in the regression and classification experiments. This would be expected as the recruiting managers have information on players over a number of years, including numerous subjective opinions on skills and fitness, performance in key games, family background, improvement over time as well as all of the draft camp data available to the neural networks. In comparison, the neural networks are limited to the draft camp data containing information on body composition, flexibility, anaerobic and aerobic power, visual tests, TAIS tests, psycho-motor tests, skill assessments and subjective assessments on strengths, weaknesses and personal attributes.

Neural network ensembles involve dividing complex problems into smaller parts and individually solving the smaller sections. Results using this technique demonstrated an improvement in performance over the single neural network approach. The experimentation conducted with ensembles in this study was preliminary in nature, however, it did demonstrate the potential for a modular approach to this type of problem.

While neural networks appear unlikely to approach the performance of the recruiting managers when used as an individual entity, the results suggest that they may have the potential to support and improve the decision making process of the RM. This is supported by the results showing that the recruiting managers alone were 44.1% correct with GOOD predictions, whereas when the recruiting managers and the NN both

agreed the percentage correct increased to 49.8%. When the RM and the NN ensemble agreed on a GOOD prediction the percentage correct increased to 52.3%. Further research is being conducted in this area.

6. Acknowledgements

The author would like to thank Stephen Wells, Kevin Sheehan, Col Hutchinson and Lee McCullagh for their cooperation and assistance in this study.

7. References

[1] Sierra, A., & Santacruz, C. Global and Local Neural Network Ensembles. *Pattern Recognition Letters*. 1998, **19**(8): 651-655.

[2] Bellerby, T., Todd, M., Kniveton, D., & Kidd, C. Rainfall Estimation from a Combination of TRMM Precipitation Radar and GOES Multispectral Satellite Imagery Through the use of an Artificial Neural Network. *Journal of Applied Meteorology*. 2000, **39**(12): 2115-2128.

[3] Smith, K. G., J. *Neural Networks in Business: Techniques and Applications.* London: IRM Press, 2003.

[4] Huang, Z., Chen, H., Hsu, C., Chen, W. & Wu, S. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems.* 2004, **37**(4): 543-558.

[5] Fieltz, L. a. S., David. Prediction of Physical Performance Using Data Mining. *Research Quarterly for Exercise and Sport*s. 2003, **74**(1): 1-24.

[6] Tschopp M., Biedert R., Seiler R., et al. Predicting success in Swiss junior elite soccer players: a multidisciplinary 4-year prospective study. *Journal of Sports Sciences.* 2003, **22**: .563 (Abstract).

[7] Wilson, R. Ranking College Football Teams: A Neural network Approach. *Interfaces*. 1995, **25**(4): 44-49.

[8] Bhandari, I., Colet,E., Parker,J., Pines,Z. Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. *Data Mining and Knowledge Discovery*. 1997, **1**: 121-125.

[9] Pyne, D., Gardner, K., Sheehan, K. and Hopkins, W. Fitness testing and career progression in AFL football. *Journal of Science and Medicine in Sport.* 2005, **8**(3): 321-332.

[10] McGee K.J. and Burkett L.N. The national football league combine: a reliable predictor of draft status. *Journal of Strength and Conditioning Research*. 2003, **17**: 6-11.

[11] Fausett, L. V. *Fundamentals of Neural Networks: Architectures, Algorithms and Applications.* Eaglewood Cliffs: Prentice Hall, 1994.

[12] Rumelhart, D. E., & McLelland, J.L. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge: MIT Press, 1986.

[13] McClelland, J.L., & Rummelhart, D.E. *Explorations in Parallel Distributed Processing*. London: MIT Press, 1991.

[14] Auda, G., & Kamel, M. Modular Neural Network Classifiers – A Comparitive Study. *Journal of Intelligent & Robotic Systems*. 1998, **21**(2): 117-129.

[15] Busson, P., Nobrega, R., & Varela, J. Modular Neural Networks for On-Line Event Classification in High Energy Physics. *Nuclear Instruments & Methods in Physics Research Section A-Accelerators Spectrometers Detectors & Associated Equipment.* 1998, **410**(2): 273-283.

[16] Parmanto, B., Munro, P.W., & Doyle, R.D. Improving Committee Diagnosis with Resampling Techniques. *Proceedings of the 1995 Conference in Advances in Neural Information Processing Systems.* 1995, **8**: 882-888.

[17] Dietterich, T. G. Machine Learning Research Four Current Directions. *AI Magazine.* 1997, **18**(4): 97-136.

[18] McCullagh, J. & Whitfort, T. An Investigation into the Application of Artificial Intelligence Techniques to the Player Selection Process at the AFL National Draft. *The 9th Australasian Conference (Tweed Heads, Sep 1-3, 2008)on Mathematics and Computers in Sport (MathSport08).* 2008, pp. 145-150.