This article was downloaded by: [Shanghai Institute of Physical Education]
On: 28 October 2013, At: 01:16
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Journal of Sports Sciences
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rjsp20

Applying decision tree induction for identification of important attributes in one-versus-one player interactions: A hockey exemplar

Stuart Morgan [a] , Morgan David Williams [b] & Chris Barnes [a]

[a] Australian Institute of Sport , Performance Research , Canberra , Australia

[b] Australian Catholic University , School of Exercise Science , Melbourne , Australia

Published online: 15 Feb 2013.

To cite this article: Stuart Morgan , Morgan David Williams & Chris Barnes (2013) Applying decision tree induction for identification of important attributes in one-versus-one player interactions: A hockey exemplar, Journal of Sports Sciences, 31:10, 1031-1037, DOI: 10.1080/02640414.2013.770906

To link to this article: http://dx.doi.org/10.1080/02640414.2013.770906

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Journal of Sports Sciences, 2013
Vol. 31, No. 10, 1031–1037, http://dx.doi.org/10.1080/02640414.2013.770906

Routledge
Taylor & Francis Group

Applying decision tree induction for identification of important attributes in one-versus-one player interactions: A hockey exemplar

STUART MORGAN[1], MORGAN DAVID WILLIAMS[2], & CHRIS BARNES[1]

[1]*Australian Institute of Sport, Performance Research, Canberra, Australia, and* [2]*Australian Catholic University, School of Exercise Science, Melbourne, Australia*

(*Accepted 24 January 2013*)

Downloaded by [Shanghai Institute of Physical Education] at 01:16 28 October 2013

Abstract

Decision tree induction is a novel approach to exploring attacker-defender interactions in many sports. In this study hockey was chosen as an example to illustrate the potential use of decision tree inductions for the purpose of identifying and communicating characteristics that drive the outcome. Elite female players performed one-versus-one contests ($n = 75$) over two sessions. Each contest outcome was classified as either a win or loss. Position data were acquired using radio-tracking devices, and movement-based derivatives were calculated for two time epochs (5 to 2.5 seconds, and 2.5 to zero seconds before the outcome occurred). A decision tree model was trained using these attributes from the first session data, which predicted that when the attacker was moving at $\geq 0.5 \text{ m} \cdot \text{s}^{-1}$ faster than the defender during the early epoch, the probability of an attacker's win was 1.00. Conversely, when the speed difference at that time was below this threshold the probability of a loss was 0.78. Secondary attributes included defender speed in the lateral direction during the early epoch, and angle of attack (i.e., angle between the respective velocity vectors of the attacker and defender) during the late epoch. The model was then used to predict outcomes of one-versus-one contests from the second session (accuracy = 0.643; area under the Receiver Operating Characteristic (ROC) curve = 0.712). Moreover, decision trees provide an intuitive framework for relating spatial-temporal concepts to coaches, and the suitability of decision trees for analysing the features of one-versus-one exchanges are discussed.

Keywords: *decision tree induction, hockey, attacker-defender dyads*

Introduction

With the advent of global positioning system (GPS), as well as emerging video-based and radio frequency (RF) technology, position tracking is now ubiquitous in elite-level invasion game sports, and the challenge for sports scientists is to draw meaning and understanding from those data. The truism that modern life is "data rich and information poor" may apply in this domain, however, techniques in data mining and statistical modelling can attach meaning and provide new insights to large sets of player tracking and position data. A feature of data mining techniques is their capacity to optimise the fit between a statistical model and the real world of data. Many techniques have already been explored in the sports domain including neural networks (e.g. Loeffelholz, Bednar, & Bauer, 2009; McGarry & Perl, 2004; Passos, Araujo, Davids, Gouveia, & Serpa, 2006), cluster analysis (Bauer & Schöllhorn, 1997; Chen, Homma, Jin, & Yan, 2007), Bayesian networks (Van Calster, Smits, & Van Huffel, 2008) and other machine learning techniques (e.g. Panjan, Sarabon, & Filipcic, 2010). In this paper we explore the potential of decision support trees for the classification of data, and to convey meaningful information to coaches.

Decision tree induction establishes a hierarchical solution to classification problems, where a set of rules is derived from the interaction between attributes in a data set. Each branch in a tree is a rule defined by splitting the values of an attribute in a way that can discriminate between states of a binary dependent variable. Each split therefore creates two new branches, which may later split in their own way to further differentiate the dependent variable. Hence, decision support trees may be described as recursive-partitioning models. The model splitting continues until such time that the model's explanatory power (on a training data set) is not further improved by additional splits (or the subsets so created are too small to be subdivided). Decision trees have been used effectively in a range of areas including medicine, financial analysis and astronomy, with the notable advantage of being suitable for highly

Correspondence: Stuart Morgan, Australian Institute of Sport, Performance Research, Canberra, Australia. E-mail: Stuart.Morgan@ausport.gov.au

© 2013 Crown copyright

Downloaded by [Shanghai Institute of Physical Education] at 01:16 28 October 2013

dimensioned data, highly scalable to large datasets, and insensitive to missing data (Han, Kamber, & Pei, 2012). Further, a significant advantage of decision tree induction is the ease of use and intuitive nature of the analysis. On the other hand the main disadvantage in decision tree induction is the propensity for over-fitting, which can make the trained model too specifically tuned to the features of the training data (Rokach & Maimon, 2008). Caution should be exercised when training a decision tree model that over-fitting is avoided.

Position data provide a particularly useful insight into the interactions between players in game sports. Indeed studies of the movement of players both in the context of within-group player dyads (Bourbousson, Seve, & McGarry, 2010a; Lames, 2006), and of group dynamics within and between teams (Bourbousson, Seve, & McGarry, 2010b; Cordes, Siegle, Stöckl, & Lames, 2010) have been especially insightful. These papers have highlighted the co-dependence of movement within and between groups, and begun to shed light on the so-called "control parameters" that are of significance in dynamical systems approaches (see McGarry, Anderson, Wallace, Hughes, & Franks, 2002).

While detailed analyses of one-versus-one game play in particular are relatively sparse, two examples warrant mention. Using phase plots of interpersonal player distance (x-axis) and relative velocity (y-axis) Passos et al. (2008) studied attacker-defender interactions in rugby. In this study, three categorical outcomes were identified for the one-versus-one interaction: a) *effective tackle*, b) *tackle with the attacker passing the defender*, and c) *clean try*. The analysis focussed on the "*critical moment*" obtained when the attacker decided to move forward and attempted to pass the defender. The authors reported random fluctuations at interpersonal distances greater than 4 m, which were described as exploratory periods where the attacker probed and sought a possible opening. Where interpersonal distance was less than 4 m, the attacker-defender dyad (driven by relative velocity) evolved to a new state: if relative velocity increased, a try was scored, but if relative velocity decreased below 2 m $\cdot$ s^{-1}, a tackle was made whether successful or unsuccessful. The authors conceded that due to the unique constraints of rugby and the age of the participants (10–12 years), generalisation to other sports and expert participants was not possible. Nevertheless, the paper introduced a rule-based description of attacker-defender interactions, which is a useful starting point for us. Employing a probabilistic model (such as decision tree induction), describing the likelihood that an outcome will result from any current state, may enhance the understanding of player interactions. We will describe in this paper how decision support trees can add such insight.

A recent study in basketball explored lateral and longitudinal movement relations between player dyads (Bourbousson et al., 2010a). In that paper the authors used relative phase analysis to explore the coordination of movement between intra- and inter-team player dyads. Bourbousson et al. (2010a) reported systematic in-phase movement for both intra- and inter-player dyads in the longitudinal direction, and also some evidence of in-phase movement for dyads in the lateral direction. In an accompanying study, Bourbousson et al. (2010b) also examined the relative phase of the team-centroids for player positions. Predictably, they also reported that the centroids of the two teams were in-phase, especially in the longitudinal direction. While the purpose of those studies was evidently to explore the spatial coupling between teams and individuals as a function of relative phase, here, as may be the case relating to the work by Passos et al. (2008), other analytical tools can provide further insight into the spatial-temporal relationships between players by the hierarchical classification of attributes that may be derived from time and position.

The objective of our study is to demonstrate the utility of decision tree induction for both understanding and communicating the features of one-versus-one exchanges in hockey. We will contend that decision trees, notwithstanding their explanatory power, also provide an intuitive framework for relating spatial-temporal concepts, and information on how best to deal with one-versus-one scenarios to coaches. Further, we will present the results of the decision tree analysis to discuss the salient features of attacker-defender interactions during one-versus-one scenarios in elite female hockey players.

Method

The Australian Institute of Sport (AIS) Human Ethics Research Committee approved the study. All participants provided informed consent and were performing regular training unrestricted by injury. Data collection was conducted in two sessions, on consecutive days, at the same location. A cohort of eight elite female hockey players volunteered for this study, four of which participated in the first session and four in the second. Regulation hockey sticks and balls were used and each player wore regular competition clothing and footwear. Each player was fitted with a Wireless Ad-hoc System for Positioning (WASP) 25 Hz RF tracking device, for which spatial accuracy has been reported at ± 0.24 m (Hedley et al., 2010). A tracking device was placed on each player's back between her shoulder blades, and was secured to the player by a custom-made elastic harness. Ten transmitting nodes were located at surveyed locations within the stadium to enable

Downloaded by [Shanghai Institute of Physical Education] at 01:16 28 October 2013

tracking coverage of the playing area. The playing area was also monitored using a standard definition digital video camera located in an elevated position from behind the position of the defensive player. Following each recording session, the raw position data and the matching video footage were downloaded directly to a hard drive on a laptop computer using the accompanying software.

All one-versus-one game samples were performed within a 10 m × 10 m area on a water-based synthetic turf pitch, which is the standard for international competition. A section of the regular white sideline was marked for the defensive end of the playing zone, and marker cones were positioned 10 m apart to indicate the three other sides of the playing area.

Before the start of each trial, the attacker and defender stood at their respective ends of the playing area, and a coach initiated play by passing the ball along the ground to the attacker. The rules of hockey were obeyed throughout all sessions and refereed by a senior member of the coaching staff. Once the attacker had received the ball in the form of a pass from the coach, the trial was "live". The attacker's objective was to advance the ball past the defender to the opposite side of the playing area without losing possession. The defender's objective was to prevent the ball reaching the end of the playing area indicated by the white sideline. The outcome for each sample was classified as either win or loss by the criteria that the attacker did (or didn't) successfully move past the defender and across the goal line while in reasonable control of the ball as judged by the coach. A loss occurred if the defender was successful in dispossessing the attacker using a legal tackle. Those samples that could not be identified as having a clear outcome (for instance if either player infringed the game rules, or the ball went outside the lateral boundaries of the playing area) were excluded from the analysis. For simplicity, and in view of the aim to describe the decision tree approach in a manageable context, this analysis was limited to the binary win/loss outcomes. Many alternate outcomes could be equally considered in a more complex model.

The players rotated their roles between attacker and defender to ensure greater diversity of playing styles. The trials were self-paced, and regular rest/drink breaks were taken to avoid fatigue. The mean duration of each trial was 6.1 seconds (± 1.6 seconds), and the mean duration of the time between trials was 40.4 seconds (± 19.2 seconds). Since the participants alternated between trials, the average rest time between each trial for a single participant was 86.9 seconds (work/rest ratio = 1:14.2).

The raw positional data and the video footage were synchronised using the corresponding video time stamp. A *Moment of Outcome* (MO) was identified in each passage of play, and defined as the point in time where either the attacking player eliminated the defender by moving past the defender in the longitudinal direction (resulting in a *win*), or the defending player dispossessed the attacker with a legal tackle (resulting in a *loss*). The subsequent temporal attributes were all calculated as a time value in relation to the MO.

All raw position data were collated and analysed using JMP 8.0 (SAS Institute Inc.). A number of parameters were derived from position and time to explore the possible candidate features for a model of one-versus-one game play. The parameters included all first order derivatives of position, including speed, direction of movement and between-player distance. Attacker (*v attack)* and defender (*v defend)* speeds were calculated as the derivative of the respective displacements in position for each player as a function of time. These values are reported as metres per second. Speed difference (*v diff)* was determined by subtracting the defender speed from the attacker speed. Attackers' lateral speed (v_x *attack*), defenders' lateral speed (v_x *defend*), attackers' longitudinal speed (v_y *attack)* and defenders' longitudinal speed (v_y *defend)* were each calculated in the same way. Lateral (v_x *diff*) and longitudinal speed differences (v_y *diff*) were then determined by subtracting the defender speed from the attacker speed in the respective direction. The Euclidean distance between the attacker and the defender at each moment *(d)* was calculated.

The angle of attack θ was calculated from the dot product of the attacker and defender velocity vectors, and can be derived as follows:

$$\theta = cos^{-1}\left(\frac{V_{attack}.V_{defend}}{||V_{attack}||.||V_{defend}||}\right) \tag{1}$$

Next, angular velocity (rate of change in the angle of attack) was calculated as follows:

$$\omega = \frac{d\theta}{dt} \tag{2}$$

Finally for the purpose of classification, mean values for each of these candidates were aggregated into two arbitrary epochs from 5 seconds before the MO to 2.5 seconds before the MO, and from 2.5 seconds before the MO to the MO itself. It would be feasible to divide the position data into smaller, one second, time epochs. Such an approach has several potential drawbacks, including the increased risk of over-fitting, and decreasing the ease of interpretation in the final model. Further research could explore the possible benefits (and costs) of smaller epochs in the context of specific player interactions. The attribute set included in the decision tree model was as follows:

Downloaded by [Shanghai Institute of Physical Education] at 01:16 28 October 2013

Mean speed difference: $\overline{v}\,diff_{5to2.5}$, $\overline{v}\,diff_{2.5to0}$

Mean lateral speeds: $\overline{v}_x\ attack_{5to2.5}$, $\overline{v}_x\ attack_{2.5to0}$, $\overline{v}_x\ defend_{5to2.5}$, $\overline{v}_x\ defend_{2.5to0}$

Mean longitudinal speeds: $\overline{v}_y attack_{5to2.5}$, $\overline{v}_y\ attack_{2.5to0}$, $\overline{v}_y\ defend_{5to2.5}$, $\overline{v}_y\ defend_{2.5to0}$

Mean lateral speed difference: $\overline{v}_x\ diff_{5to2.5}$, $\overline{v}_x\ diff_{2.5to0}$

Mean longitudinal speed difference: $\overline{v}_y\ diff_{5to2.5}$, $\overline{v}_y\ diff_{2.5to0}$

Mean Euclidean distance between players: $\overline{d}_{5to2.5}$, $\overline{d}_{2.5to0}$

Mean angle of attack: $\overline{\theta}_{5to2.5}, \overline{\theta}_{2.5to0}$

Mean angular velocity of attack: $\overline{\varpi}_{5to2.5}, \overline{\varpi}_{2.5to0}$

A dataset of 75 trials were recorded across Session 1 and Session 2. To avoid over-fitting, it is normal practice to divide a data set into two sections, for the testing, and subsequent training of the model. In this instance the valid trials recorded in Session 1 ($n = 33$) were used to train the model, while the trials recorded in Session 2 ($n = 42$) with a different cohort of players were subsequently used to test the model. Each of the attributes described above were included as factors, and the binary outcome of the trials (win or loss) was included as the dependent response.

Results

The features of the decision tree model are presented in Figure 1 below. The first node includes all trials in the training dataset ($n = 33$), of which 21 resulted in a loss, and 12 resulted in a win. The first split in the model occurred for the attribute $\overline{v}\ diff_{5to2.5}$ such that for trials where $\overline{v}\ diff_{5to2.5}$ was equal or greater than 0.5 $\text{m} \cdot \text{s}^{-1}$ ($n = 6$), the probability of a win was 1.00. Conversely, where $\overline{v}\ diff_{5to2.5}$ was less than 0.5 $\text{m} \cdot \text{s}^{-1}$ ($n = 27$), the probability of a win was 0.22. Splitting the latter node using the $\overline{v}_x\ defend_{5to2.5}$ attribute made further improvements. Where $\overline{v}_x\ defend_{5to2.5}$ was less than 1.4 $\text{m} \cdot \text{s}^{-1}$ ($n = 13$) the probability of a win was 0.00, and where $\overline{v}_x\ defend_{5to2.5}$ was equal or greater than 1.4 $\text{m} \cdot \text{s}^{-1}$ ($n = 14$) the probability of a win was 0.43. The model was further split on that node for the attribute $\overline{\theta}_{2.5to0}$. On this final node, where $\overline{\theta}_{2.5to0}$ was less than 17° ($n = 7$) the probability of a win was 0.71, and where $\overline{\theta}_{2.5to0}$ was equal or greater than 17° ($n = 7$) the probability of a win was 0.14. A descriptive set of IF-THEN rules can then be extracted from the decision tree. The rules are further described in the form of IF-THEN rules in Table I.

The overall model was evaluated both in terms of its fitting of the training set data from Session 1, and in terms of its predictive value for the testing data set from Session 2. The model demonstrated a good fit with the training data set, and the R^2 coefficient was 0.736 indicating that the model explained nearly three-quarters of the variation in the outcome of trials in Session 1. The *sensitivity* of the model relates to the proportion of true positives to all actual positives (wins), and is calculated by dividing the number of true positives (correctly predicted wins) by the sum of true positives and false negatives (all actual wins). The sensitivity of the trained model for the Session 1 data set was 0.917. Similarly, *specificity* relates to the model's capacity to detect losses as such (true negatives divided by the sum of true negatives and false positives). The specificity of the trained model for the Session 1 data was 0.952.

In tuning a model that is highly sensitive to the detection of true cases (in our model, wins), sensitivity can often be at the cost of specificity, where false positives result from eagerness to detect all wins. The Receiver Operating Characteristic (ROC) presents a graphic indicator of sensitivity compared with specificity, and the area under the curve provides a summary statistic of the curve. It is formed using the probability of the predicted outcome for each trial. Trials are ordered by rank such that the trial most likely to be a win appears at the top of the list. The ordered set of sensitivity and specificity values are shown in the graphic with the former on the y-axis and the latter on the x-axis. In a binary model such as this, the ROC curve shows the trade-off between sensitivity to detecting actual wins, and the specificity of not mistaking losses for wins (for a more detailed explanation of ROC curves refer to Bradley, 1997; Fawcett, 2006). The ROC for Session 1 is presented in Figure 2, and the area under the ROC is 0.960, which provides a discrete measure of the fit between the model and the data, indicating a very close agreement between the model and the training data set. While it may be possible to further refine the attributes included in the model for an even closer fit to the training data, it is possible that such a model would become over-fitted to outliers in the training set. The objective in tuning a decision tree model is to establish a valid model that demonstrates the capacity to accurately classify independent data sets.

The model presented above was applied to the independent data set from Session 2. The predictive capacity of the model is measured in several ways. In Table II, the normalised confusion matrix shows the relationship between the model predictions and the actual trial outcomes as a proportion of all trials in Session 2. Table II indicates that 0.143 of all trials were correctly classified wins, and half of all trials (0.500) were correctly classified losses. Hence the overall accuracy of the model for the testing data is the sum of true positives and true negatives, which was 0.643. Further, sensitivity in the testing data was 0.667, indicating the proportion of all wins that were correctly classified, and the specificity was 0.636, indicating the proportion of all losses that were correctly classified as such. Finally, the ROC for the

Downloaded by [Shanghai Institute of Physical Education] at 01:16 28 October 2013

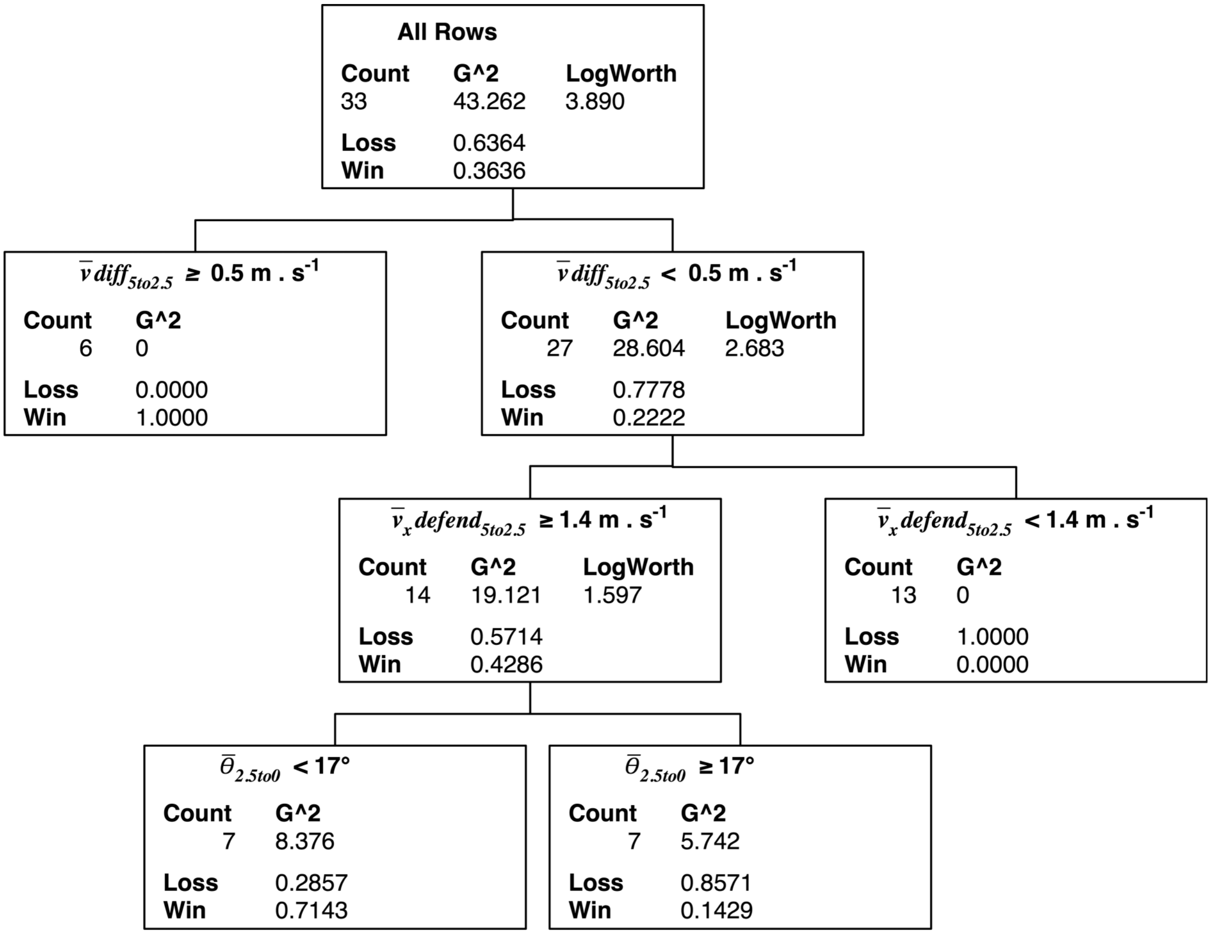
All Rows
Count 33 G^2 43.262 LogWorth 3.890
Loss 0.6364
Win 0.3636
$\bar{v}\,diff_{5to2.5}$ ≥ 0.5 m . s^{-1}
Count 6 G^2 0
Loss 0.0000
Win 1.0000
$\bar{v}\,diff_{5to2.5}$ < 0.5 m . s^{-1}
Count 27 G^2 28.604 LogWorth 2.683
Loss 0.7778
Win 0.2222
$\bar{v}_x defend_{5to2.5}$ ≥ 1.4 m . s^{-1}
Count 14 G^2 19.121 LogWorth 1.597
Loss 0.5714
Win 0.4286
$\bar{v}_x defend_{5to2.5}$ < 1.4 m . s^{-1}
Count 13 G^2 0
Loss 1.0000
Win 0.0000
$\bar{\theta}_{2.5to0}$ < 17°
Count 7 G^2 8.376
Loss 0.2857
Win 0.7143
$\bar{\theta}_{2.5to0}$ ≥ 17°
Count 7 G^2 5.742
Loss 0.8571
Win 0.1429

Figure 1. Decision tree nodes for trained model.

Table I. Binary predictions for decision tree model as IF-THEN rules.

Rule 1	**IF** $\bar{v}\ diff_{5to2.5} \geq 0.5\ \mathrm{m} \cdot \mathrm{s}^{-1}$	THEN ***WIN***
Rule 2	**IF** $\bar{v}\ diff_{5to2.5} \geq 0.5\ \mathrm{m} \cdot \mathrm{s}^{-1}$ **AND** $\bar{v}_x defend_{5to2.5} < 1.4\ \mathrm{m} \cdot \mathrm{s}^{-1}$	THEN ***LOSS***
Rule 3	**IF** $\bar{v}\ diff_{5to2.5} \geq 0.5\ \mathrm{m} \cdot \mathrm{s}^{-1}$ **AND** $\bar{v}_x defend_{5to2.5} \geq 1.4\ \mathrm{m} \cdot \mathrm{s}^{-1}$ **AND** $^{-1}\bar{\theta}_{2.5to0} < 17\ °$	THEN ***WIN***
Rule 4	**IF** $\bar{v}\ diff_{5to2.5} \geq 0.5\ \mathrm{m} \cdot \mathrm{s}^{-1}$ **AND** $\bar{v}_x defend_{5to2.5} \geq 1.4\ \mathrm{m} \cdot \mathrm{s}^{-1}$ **AND** $^{-1}\bar{\theta}_{2.5to0} \geq 17\ °$	THEN ***LOSS***

model applied to the testing set is presented in Figure 3. The area under the ROC is 0.712, which indicates a moderate agreement between the trained model and the testing data set.

Discussion

The aims of this study were to demonstrate that decision tree induction can be used to describe a basic model of one-versus-one game play in women's hockey, and to resolve important predictive attributes in the outcome of game sports scenarios. Decision tree partitioning requires no *a-priori* knowledge about model parameters, and therefore no prior selection of parameters is required. We compiled a comprehensive set of movement and position related attributes in the composition of our model, although further research could explore other parameters also. In descriptive terms our model suggests that speed difference is the single most important feature determining the outcome of a one-versus-one exchange in hockey. Where the attacker is moving more quickly than the defender in the early epoch by greater than the critical threshold difference ($\bar{v}\ diff_{5to2.5} \geq 0.5\ \mathrm{m} \cdot \mathrm{s}^{-1}$), then there is a 100% prediction that the attacker will win the exchange. Where speed difference is less than this critical threshold, there is a 77.8% prediction probability that the attacker will be defeated. In fact, if we were to prune the decision tree to this point, then the area under the ROC curve would still be 0.75 for the training data, and 0.656 for the testing data. The model at this level indicates clearly that attackers generate success by moving quickly (more so than the defender) and early in a one-versus-one exchange.

Downloaded by [Shanghai Institute of Physical Education] at 01:16 28 October 2013

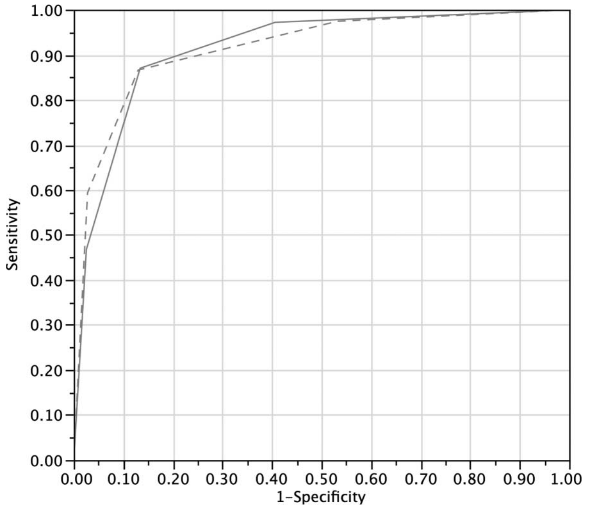

Figure 2. ROC curve for training data set (Session 1). The solid curve shows the model's performance as a trade-off between sensitivity and specificity in predicting attacking wins in the training data. The dashed curve shows the same information for attacking losses. The area under the curve is 0.960.

Table II. Normalised confusion matrix showing the relationship between model predictions and actual trial outcomes as proportions of all trials in Session 2.

	Pred. Win	Pred. Loss
Actual Win	0.143	0.071
Actual Loss	0.286	0.500

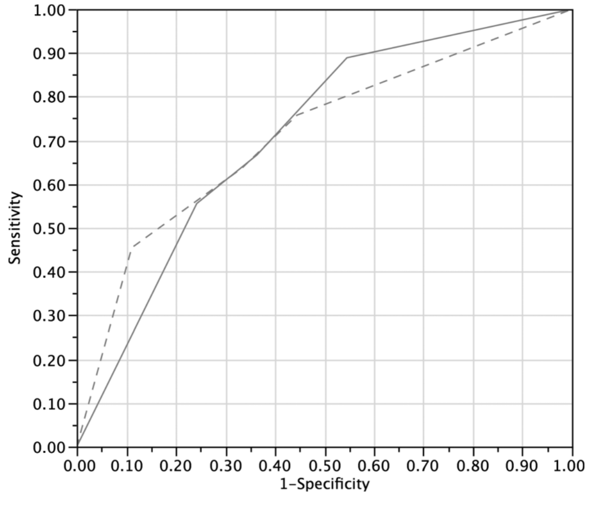

Figure 3. ROC curve for testing data set (Session 2). The solid curve shows the model's performance as a trade-off between sensitivity and specificity in predicting attacking wins in the testing data. The dashed curve shows the same information for attacking losses. The area under the curve is 0.714.

Following the model further, where speed difference between the attacker and defender is less than this critical threshold, the defender's lateral speed becomes the next-most important feature of the exchange. Our model suggests that the odds are already in favour of the defender if speed difference is small, but the attacker's odds can improve if the defender is moving laterally at a speed more than 1.4 m $\cdot$ s^{-1} in the early epoch. However, if the defender is moving laterally at less than this speed then the model predicts a 100% chance the defender will win the exchange. In operational terms then, the coach's advice to an attacker might be that if you cannot move at a speed significantly faster than the defender, then you at least need to manipulate the defender in a way that has them moving rapidly sideways.

The final branch in the tree refers to the angle of attack in the final 2.5 seconds of the exchange ($\overline{\theta}_{2.5to0}$). Where the attacker and defender run directly towards each other from exactly opposite directions then $\overline{\theta}_{2.5to0}$ is equal to 180°. Where they run in exactly the same direction then $\overline{\theta}_{2.5to0}$ is equal to 0°. In our model, the $\overline{\theta}_{2.5to0}$ attribute differentiates the likely outcome where the preceding rules $\overline{v}\ diff_{5to2.5} < 0.5\ \mathrm{m} \cdot \mathrm{s}^{-1}$, and $\overline{v}_x\ defend_{5to2.5} \geq 1.4\ \mathrm{m} \cdot \mathrm{s}^{-1}$ are true. In those cases, where $\overline{\theta}_{2.5to0} < 17$ °, and the attacker and defender are running closer to parallel, then the model predicts a probable win (71.4%) for the attacker. Alternately the probability of a win is just 14.3% where $\overline{\theta}_{2.5to0}$ is greater than that threshold.

The angle of attack parameters in the late epoch do not convey the specific direction of movement. We may still infer that the attacker is moving in a generally forwards direction towards the goal line, and the defender is retreating with the attacker. It is common practice in hockey to defend in this way, and for the defender to move in a parallel direction to the attacker and patiently wait for an opportunistic moment to attempt a dispossession. Where the angle of attack is less, then the attacker and defender are moving in parallel lines, more or less directly towards the goal line, in which case the contest may become a virtual foot race. The odds in our model suggest that this scenario favours the attacker. Alternately, where the angle of attack is greater, we might assume that the attacker is taking a more indirect route towards the goal line. Alternately it is possible that the defender is changing direction during the exchange, although the more intuitive interpretation of this relationship is that the defender would follow the movements of that attacker. Whichever is the case, the odds in this scenario are in favour of the defender.

Therefore, a coach could encapsulate the entire model in operational terms as advice to attackers in one-on-one encounters: move fast and early, or at the very least shift the defender early and laterally, and then run a direct line to the goal area. It is an appealing feature of the rule-based decision tree format that complex positional and movement information can be represented in a convenient operational context such as this.

Downloaded by [Shanghai Institute of Physical Education] at 01:16 28 October 2013

Overall we have presented a model that is based on movement and position-based parameters. Further research should consider other possible attributes that may increase the predictive validity of the model with independent data sets. Certainly ball position may be a relevant feature in determining the outcome of one-versus-one game play. Generally in hockey, the ball is carried on the right side of the body (as there are no left-handed hockey sticks), which allows the attacking player to protect and control the ball with greater confidence. Nevertheless, elite hockey players can position the ball on either side of their body while running, and the transition of the ball from one side to another can be used to manipulate the position of the defender. Equally, elite players have the ability to shift the ball around their body very quickly, and they are well practised at creating deception about where they intend to move the ball. This is often called "selling the dummy", and is stock in trade for skilful attackers trying to manipulate the position of the defender. A more detailed model of the one-versus-one scenario could attempt to include ball position factors in the analysis, although this would require a novel method for tracking the hockey ball.

In this study we trained a decision tree using elite hockey players, and tested the model using a separate elite cohort. The model is predictive of the outcome of confined one-versus-one plays in that context, but it is unclear how well that model would predict one-versus-one outcomes in more generalised contexts such as normal matches. In open game play the attacking ball carrier has many additional degrees of freedom, where they might attempt to eliminate a single defender, or otherwise pass the ball to another player, or deliberately engage with several defenders to create attacking opportunities for other teammates. Although our model describes a single scenario within a larger game, our aim is to provide researchers with a methodological framework to address more complex aspects of team invasion sports such as hockey. A significant advantage in decision tree induction is scalability to large data sets and many attributes. Future work will explore open game play taking into account a much larger field of candidate attributes relating to multiple player positions.

Conclusions

Modern techniques in data collection such as RF and video-based tracking, and GPS have introduced many new sources of data in sport. The next major challenge for sports scientists is to develop methods to resolve insight from large data sets, and to find ways to convert complex and multi-dimensional data into meaningful recommendations for coaches. Here we have presented a simple model of the one-versus-one game scenario in hockey using decision tree induction. Although the specific features of the model may not be widely applicable to other contexts, our aim has been to demonstrate the predictive and descriptive features of this method in a simplified setting.

References

Bauer, H. U., & Schöllhorn, W. (1997). Self-organizing maps for the analysis of complex movement patterns. *Neural Processing Letters, 5*(3), 193–199.

Bourbousson, J., Seve, C., & McGarry, T. (2010a). Space–time coordination dynamics in basketball: Part 1. Intra- and inter-couplings among player dyads. *Journal of Sports Sciences, 28*(3), 339–347.

Bourbousson, J., Seve, C., & McGarry, T. (2010b). Space–time coordination dynamics in basketball: Part 2. The interaction between the two teams. *Journal of Sports Sciences, 28*(3), 349–358.

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.

Chen, I., Homma, H., Jin, C., & Yan, H. (2007). Clustering and display of elite swimmers' race patterns across various comparable criteria at the same time. *International Journal of Sports Science and Engineering, 1*(2), 129–136.

Cordes, O., Siegle, M., Stöckl, M., & Lames, M. (2010). Coupling of players and teams in soccer analyzed by Relative Phase. In R. Stretch (Ed.), *The second world conference on science and soccer. Conference program* (S. 125). Port Elizabeth, South Africa: Nelson Mandela Metropolitan University (NMMU).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Han, J., Kamber, M., & Pei. J. (2012). *Data Mining: Concepts and techniques* (3rd ed.). San Francisco, CA: Morgan Kaufmann.

Hedley, M., Mackintosh, C., Shuttleworth, R., Humphrey, D., Sathyan, T., & Ho, P. (2010). *Wireless tracking system for sports training indoors and outdoors*. Paper presented at the 8th Conference of the International Sports Engineering Association (ISEA), Vienna, Austria.

Lames, M. (2006). Modeling the interaction in game sports - relative phase and moving correlations. *Journal of Sports Science and Medicine, 5*(4), 556–560.

Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports, 5*(1), 1–15.

McGarry, T., Anderson, D. I., Wallace, S. A., Hughes, M. D., & Franks, I. M. (2002). Sport competition as a dynamical self-organising system. *Journal of Sports Sciences, 20*(10), 771–781.

McGarry, T., & Perl, J. (2004). Models of sports contests: Markov processes, dynamical systems and neural networks. In M. D. Hughes & I. M. Franks (Eds.), *Notational analysis of sport* (2nd ed.) (pp. 227–242). London: Routledge.

Panjan, A., Sarabon, N., & Filipcic, A. (2010). Prediction of the successfulness of tennis players with machine learning methods. *Kinesiology, 42*(1), 98–106.

Passos, P., Araujo, D., Davids, K., Gouveia, L., & Serpa, S. (2006). Interpersonal dynamics in sport: The role of artificial neural networks and 3-D analysis. *Behavior Research Methods, 38*(4), 683–691.

Passos, P., Duarte, A., Davids, K., Gouveia, L., Milho, J., & Serpa, S. (2008). Information-governing dynamics of attacker-defender interactions in youth rugby union. *Journal of Sports Sciences, 26*(13), 1424–1429.

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: Theory and applications*. Singapore: World Scientific.

Van Calster, B., Smits, T., & Van Huffel, S. (2008). The curse of scoreless draws in soccer: The relationship with a team's offensive, defensive, and overall performance. *Journal of Quantitative Analysis in Sports, 4*(1), Article 4, 1–22.