Measurement in Physical Education and Exercise Science, 17: 171–186, 2013
Copyright © Taylor & Francis Group, LLC
ISSN: 1091-367X print / 1532-7841 online
DOI: 10.1080/1091367X.2013.805137

Data Mining in Elite Sports: A Review and a Framework

Bahadorreza Ofoghi
Institute of Sport, Exercise, and Active Living, Victoria University
Melbourne, Victoria, Australia
Victorian Institute of Sport
Albert Park, Victoria, Australia

John Zeleznikow and Clare MacMahon
Institute of Sport, Exercise, and Active Living, and School of Management and Information Systems, Victoria University
Melbourne, Victoria, Australia

Markus Raab
Institute of Psychology, German Sport University Cologne
Cologne, Germany

Sophisticated data analytical methods such as data mining, where the focus is upon exploration and developing new insights, are becoming increasingly useful tools in analysing elite sports performance data and supporting decision making that is crucial to gaining success. In this article, we investigate the different data mining demands of elite sports with respect to a number of features that describe sport competitions. The aim is to more structurally connect the sports and data mining domains through: (a) describing a framework for categorizing elite sports, and (b) understanding the analytical demands of different performance analysis problems. Therefore, we review different aspects such as sport categories and performance analysis requirements that influence each stage in sports data mining. We also present a model bringing together performance analysis requirements, data mining methods, data mining techniques, and technique characteristics. This will assist both data scientists and sport professionals to more effectively collaborate and contribute to success in elite sport events.

Key words: data mining, elite sport, performance analysis

Strategic decisions in sports are vital for performance outcomes. These includes decisions such as how to avoid elimination in swimming preliminary races without an over-expenditure of energy that will reduce performance in subsequent races, or the use of second-string players in early rounds of hockey tournaments to avoid injury and fatigue to star players. The importance of these decisions is often matched by their complexity. For example, when we asked the Victorian Institute of Sport netball coach “what are your crucial problems?,” she interpreted and answered

Correspondence should be sent to Bahadorreza Ofoghi, Institute of Sport, Exercise, and Active Living, Victoria University, PO Box 14428, Melbourne, Victoria 8001, Australia. E-mail: bahadorreza.ofoghi@vu.edu.au

the question from several different levels: a) what events in the match have the strongest influence on the outcome or b) what events in the match influence “shooting percentage,” and “which center passes lead to shots on goal.” The complexity of these problems requires sophisticated performance analysis methods.

Performance analysis, as a means to create and analyze a valid record of athlete performances by using systematic observations, has gained importance in the last decade. It has been facilitated by advances in information technology and digital photography (Bishop, 2003). One such advanced technology is data mining, which is a branch of computer science and artificial intelligence. Data mining techniques can be used to analyze sports data, particularly elite sports performance data (Schumaker, Solieman, & Chen, 2010). To specify, we refer to elite sports performance data as it is commonly referred to, as data arising from international competitions such as world championships or Olympic games.

This review article focuses on the problems encountered in the analysis of archived elite sports performance data using data mining techniques. It addresses the apparent lack of a well-defined data analytical framework that considers the demands of different problems in elite sports and sport performance analysis requirements. Most of the previous attempts at using sophisticated data mining methods and techniques for sports performance analysis have only considered ad hoc problems that arise in specific sports, and even so, only in a limited context (e.g., finding inter-relationships between women’s 50-m breaststroke swimming performance attributes, such as start time, breakout time, and breath count, and their overall swim times; Ofoghi, Stefano, Zeleznikow, & MacMahon, 2012). We present a framework for categorizing different sports with respect to pre-specified attributes that relate to performance analysis. We also identify the analytical demands of different types of research problems in elite sports in terms of performance analysis.

In order to achieve our goals, this review systematically maps data mining methods and sports performance analysis. Specifically, we will clarify what data mining is, what data mining methods exist, and how these techniques can be gainfully employed in elite sports. We will provide a rationale for how data mining and data analysis benefit elite sport competitors. As such, this review focuses on the exploratory level of data analysis where the extraction of useful information from previous data is the major concern. This review article, therefore, leaves aside the explanatory level that explains and justifies any useful information discovered from the data.

DATA MINING: A BRIEF INTRODUCTION

In the information science domain, converting raw data into information, information into knowledge, and knowledge into wisdom forms a well-known hierarchy of data processing also known as the *wisdom hierarchy* (Rowley, 2007). Data mining is mainly concerned with the first part of the wisdom hierarchy to convert data into information. It is a problem-solving methodology that finds a logical or mathematical description of patterns and regularities in a set of data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Useful and previously unknown information can be extracted from archived or streamed data by using data mining methods. The extracted information may be in the form of prediction of events (Anagnostopoulos, Anagnostopoulos, Hadjiefthymiades, Kalousis, & Kyriakakos, 2007), finding events that co-occur and their sequence of occurrence

(Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1995), and division of data into groups of similar objects (Berkhin, 2006).

In the elite sports domain, data mining methods have generally been used to model the interrelationships of performance measures and attributes (e.g., Ofoghi et al., 2012; Wilson et al., 2001) and to also extract athlete performance patterns from previously held competitions (e.g., Chen, Homma, Jin, & Yan, 2007; Edelmann-Nusser, Hohmann, & Henneberg, 2002; Ofoghi, Zeleznikow, MacMahon, & Dwyer, 2010). These can be used in the decision-making processes to support strategic planning and athlete selection.

Commonly Used Data Mining Methods in Elite Sports

Clustering. Clustering is one form of unsupervised learning that is concerned with finding how the data are organized and summarizing/explaining key features of the data (Clausen, 2012). The result of a cluster analysis is the formation of a number of groups. The members of each group are similar to each other regarding some criteria (i.e., similarity attributes) and are most dissimilar to those of the other groups with respect to the same criteria. Table 1 summarizes the major uses of clustering for sport data analysis.

Classification. Classification is a form of supervised learning where the aim is to learn the nature of the mapping from a given input to a given output (Alpaydin, 2010). The system is repeatedly given facts about various cases, along with (known) expected outputs. It then adjusts the weights of a mapping function that can produce outputs similar to the expected results. Once the mapping function is stable (i.e., training is completed), the function can be used to predict group memberships for unseen data instances (i.e., new individual cases). Contrary to clustering, where

TABLE 1
Utilization of Clustering, Classification, and Relationship Modeling Techniques in the Sports Domain

Method	*Researcher(s)*	*Technique*	*Sport*
Clustering	Gaudreau & Blondin (2004)	ward	golf
	Ball & Best (2007)	k-means	golf
	Chen et al. (2007)	average linkage (hierarchical)	swimming
	Woolf et al. (2007)	mixed	decathlon
	Lamb et al. (2010)	self-organizingmaps	basketball
	Ofoghi et al. (2010)	k-means	track cycling
Classification	Ofoghi et al. (2010)	Naive Bayes	track cycling
	Watson (1988)	linear discriminant analysis	rugby
	Smith & Spinks (1995)	linear discriminant analysis	rowing
	Jaitner et al. (2001)	linear discriminant analysis	long jump
Relationship modeling	Wilson et al. (2001)	linear regression	swimming
	Johnson et al. (2009)	linear and polynomial regression	swimming
	Edelmann-Nusser et al. (2002)	neural networks	swimming
	Shao (2009)	neural networks	aerobics
	Kahn (2003)	neural networks	football (NFL)
Rule mining	Bhandari et al. (1997)	association rules	basketball (NBA)
	Sun et al. (2010)	association rules	table tennis

Notes. NFL = National Football League; NBA = National Basketball League.

there is no or limited prior knowledge about data groups/memberships available beforehand, in classification, the class of training data is known and that of unseen data can be predicted after training is complete. Table 1 summarizes some previous studies that have used classification techniques for sports performance analysis.

Relationship modeling. The goal of relationship modeling in a complex environment is to find a function or model that best describes the inter-relationship between predictor and dependent attributes. The main aim in this task is to find a model that fits data with the least error.

Regression analysis (Freund, Wilson, & Sa, 2006) is one of the methods that has been used for fitting a line or polynomial to data. The idea of this analysis is to relate a response variable to a vector of predictor variables (Smyth, 2002). The relationship found can then be used for prediction purposes. More sophisticated methods for relationship modeling are neural networks that can be used to find attribute inter-relationships (Edelmann-Nusser et al., 2002; Kahn, 2003; Shao, 2009). A neural network receives its name from its resemblance to a nervous system in the brain. It consists of many self-adjusting processing elements cooperating in a densely interconnected network. Each processing element generates a single output signal that is transmitted to the other processing elements. The output signal of a processing element depends on the inputs to the processing element: each input is gated by a weighting factor that determines the amount of influence that the input will have on the output. The strength of the weighting factors is adjusted autonomously by the processing element as data is processed (Stranieri & Zeleznikow, 2005). Some of the previous studies that have used relationship modeling for sports data analysis are summarized in Table 1.

Rule mining. Finding rules that represent inter-relationships between series of events (e.g., pressing in the first half of a football game and winning the game) or states (e.g., being emotionally down or being physically weak) has two main forms in data mining, namely association rule mining (Agrawal et al., 1993) and sequential pattern mining (Agrawal & Srikant, 1995). While association rule mining is only concerned with finding events or states that co-occur, sequential pattern mining is focused around the sequence of events that co-occur with a high frequency in a timestamp-ordered set of events.

These two types of rules/patterns (association rule and sequential pattern mining), when used in the sports domain, can potentially reveal series of conditions, movements, decisions, positions, or events in general, as well as their sequences in time, that may lead to certain positions, scores, or outcomes. Table 1 includes some examples of the use of rule mining as applied to the elite sports performance analysis.

The Benefits of Data Mining for Analyzing Elite Sports Data

Statistical measures have been used for most of the performance analysis work in sports, specifically for match analysis. According to Bishop (2003), “match analysis means to record aspects of team performance” (p. 1). We extend this definition to “match analysis is the analysis of important attributes, events, strategies, or patterns (and their importance) in gaining success in different contests.” An example is tracking the number of passes before a goal or the areas on the field where goals are scored from. These statistical measures range from what we consider straightforward to sophisticated analyses. To us, a straightforward analysis is usually concerned with finding direct relationships between a few predictor variables and a dependent variable. This may ignore the overall context of the data that have been collected (e.g., regression analysis of variables without considering other conditions that may have led to certain variable relationships). Examples of studies where straightforward data analysis was conducted include Pollard and Pollard (2010), Ransdell, Vener, and Huberty (2009), and Vezos et al. (2007).

In what we consider sophisticated approaches to match analysis, more in-depth analysis is carried out that goes beyond the surface features and structure of performance attributes. The aim is to find more hidden, underlying relationships between factors that may either directly or indirectly influence sports performances with respect to different target variables, such as overall rankings, finishing times, or even physiological parameters that may lead to certain performances. More sophisticated data analysis studies include those carried out by Cox and Dunn (2002), Kenny, Sprevak, Sharp, and Boreham (2005), Kline et al. (2007), Liao (2008), Vaz, Rooyen, and Sampaio (2010), and Zwols and Sierksma (2009). These studies, for instance, have shed light on circadian variation in swimming or determinants of success in decathlon. But the use of statistical analysis without an understanding of the fundamental meaning of the analyses and the contextual backgrounds of variables/values under study, such as those in the task constraints of a decathlon, can potentially be misleading due to impreciseness or over emphasis of the statistics. Impreciseness may occur if other contributing variables are missed or ignored. Avoiding this impreciseness may de-emphasize over-interpretations. Such a less-informed approach (i.e., pure statistical analysis) may also reduce the potential for generalizations to other sports (Schumaker et al., 2010). Over emphasis can be due to the effect of conceptually non-related coincidences being ignored. The utilization of data mining techniques, above more simple statistical analyses, is a way to overcome these problems. For instance, taking contextual information into data mining analyses, such as performance variations of one's own team for specific groups of opposing teams, circumvents impreciseness of mean statistics.

Data Mining Demands in Analyzing Elite Sport Performance Data

When analyzing sports performance data, there are three main attributes that interest sport scientists and professionals, namely rankings, times, and scores (RTS). These three measures, represent performance and are used for evaluating athletes in most sport events.

The existing gap between applied sport and data mining can be addressed through a mapping, as shown in Figure 1a, from the sports domain to the data mining domain. Such a mapping allows for the selection of the appropriate data analytical technique for the sports question at hand. While the sports domain involves the main rules, regulations, tactics, strategies, performances, conditions, and abilities related to specific sports, the data mining domain includes the representative performance measures, namely RTS measures.

The data preprocessing and data analysis methods that can be utilized in the data mining space can only interpret the available data in the form of the performance measures. A deeper understanding of (sports) domain knowledge as well as a better understanding of the available and appropriate data mining tools serves to minimize problems in sports performance analysis that may arise due to the lack of precision, sound approach, or validity of the analysis.

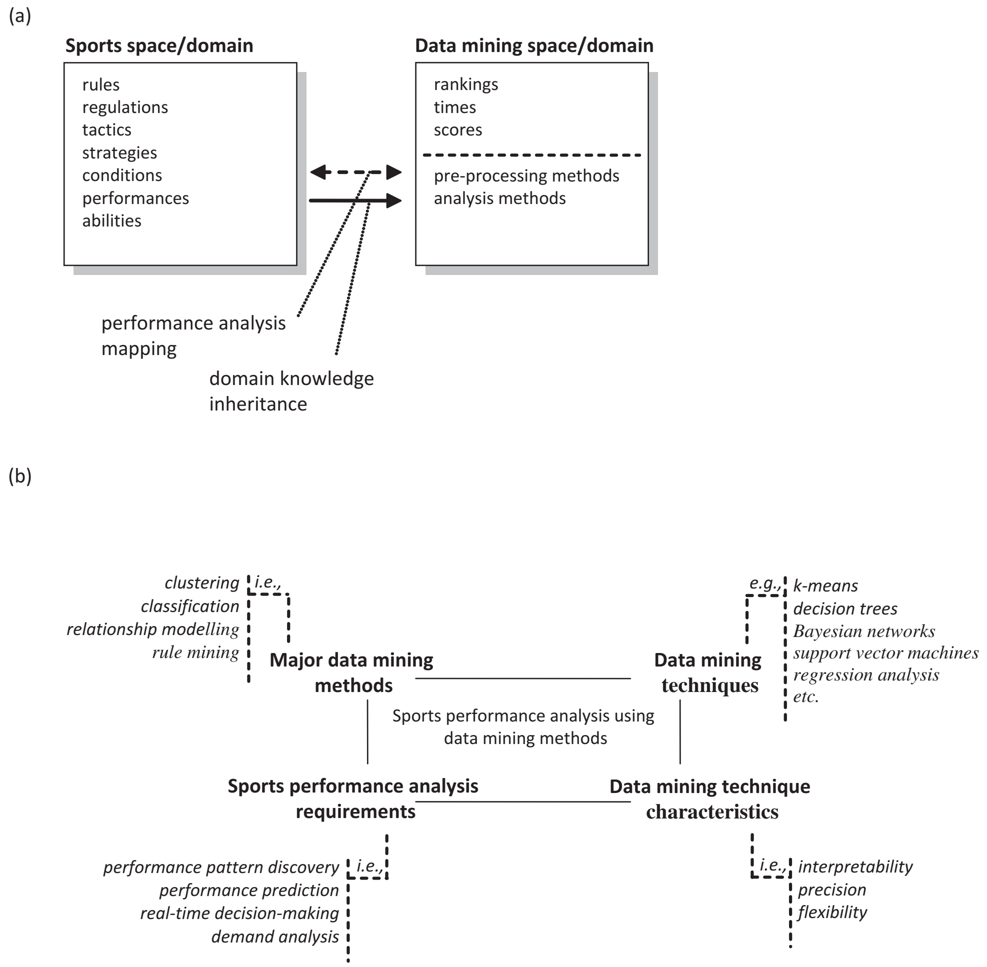

FIGURE 1 (a) The sport performance analysis scheme involving the sports domain and the data mining domain. (b) The rectangular model characterizing the data mining approach towards sports performance analysis.

The mapping that occurs from the sports domain to the data mining domain for performance analysis will serve as a starting point to combine research in sport science and computer science more effectively. Combining research between disciplines requires that the mapping has the ability to inherit required (sports) domain knowledge. This knowledge is necessary for data analysts not only to evaluate the results of their analysis but also to validate the processes that they carry out for preparing data and extracting information from this data. Knowledge about sports exists at different levels, such as formal rules (e.g., knowledge about rankings and scoring) and informal rules (e.g., ethical issues) of individual sports or the sports domain in general.

Preprocessing of Data in Sports

Sports-related data preprocessing, in the data mining space, involves preparing and sorting data for analysis. Depending on the current and required formats of the data and the research problem, preprocessing can consist of one or more of the following tasks:

- Filtering: Filtering data records into certain categories of competitions (e.g., categorizing rowing performance results into fast, medium, and slow courses).
- Format conversion: Converting data into a format which can be interpreted by specific data analytical software and tools that are used to conduct actual data analysis (e.g., converting hh:mm:ss times into collective seconds).
- Extraction: Finding new data not explicitly available based on collected data (e.g., extracting absolute and cumulative rankings of boats relative to different sectors of 2000-m rowing races based on the times of the boats; Ofoghi, Zeleznikow, & MacMahon, 2011a).
- Structural conversion: Converting parts of data into a format that allows for more precise data analytical processes (e.g., generalizing final standings ranging from 1 to the number of contestants into medal winner [positions 1–3] and non-medal winner [positions greater than 3] categories).
- Descriptive conversion: Converting specific parts of data to a format that better describes the nature of the specific sport/problem (e.g., converting absolute times to relative/differential times that show the time differences between the leader and other athletes).

While the first three types of data preprocessing are straightforward, the distinction between the last two (structural and descriptive conversion) may be more delicate and less tangible. In structural conversion, the semantics of data will not change (e.g., the rankings before and after generalization are still of the same semantics) whereas in descriptive conversion, the meaning of the new converted data will not be the same as its original form (e.g., times vs. time differences or timestamp labelled and ordered positions vs. raw positions).

The data preprocessing tasks in filtering, format conversion, and extraction are not affected by the specifics of a given sport. But structural conversion is usually dependent on: (a) the specifics of the data analytical method whereby some data analytical methods can more effectively handle nominal values compared to numerical data types (e.g., classification systems) and (b) the amount of data that is available whereby often small amounts of data pertaining to specific parameters may necessitate generalization of the values in to more coarse-grained values (e.g., generalized finishing places, such as the categories defined for the track cycling omnium by Ofoghi, Zeleznikow, MacMahon, and Dwyer [2010] that cover a greater number of data records).

Descriptive conversion is closely linked to understanding the sport and its features. The appropriate data preprocessing technique may be chosen or driven by the combination of features of a sport, such as number of events, number of athletes, duration, or winning criteria. We will address these issues individually in the next section. In the cases where there are a number of sport features to be considered, a combination of preprocessing tasks may be appropriate.

The effect of the number of events in a competition on descriptive conversion. Most single-component sports (e.g., cycling time-trial and marathon running) do not usually require or influence descriptive conversion in data preprocessing. The main reason for this is, for these sports, the predictor variables (i.e., the RTS measures) explicitly represent the nature of the sport.

These variables can therefore be utilized for modeling the underlying structure to predict the variations of the dependent variable (e.g., the final standings). For instance, 200-m freestyle swimming does not require any specific type of preprocessing of lap times to predict finishing times.

Multiple-component sports, however, may necessitate converting specific data to other forms. We argue that this depends on whether the individual events are held successively (e.g., triathlon) or independently (e.g., track cycling omnium). RTS in independent multiple-component sports usually provides enough information for modeling and predicting dependent variables (i.e., the overall times and the overall final standings). The machine learning-based analysis (clustering and classification) of the necessary abilities for winning the track cycling omnium competitions carried out by Ofoghi et al. (2010), for example, includes no further preprocessing on the rankings of riders in each individual event than generalizing the final standings into three categories. The categories are: medal winners, non-medal winners ranked between 4 and 10, and non-medal winners ranked above 10. This preprocessing is mainly due to the small amount of historical data which is available for this specific sport.

In successive multiple-component sports, in contrast, raw RTS measures may be misleading in the performance analysis task. The triathlon is an example of a successive multiple-event competition where the raw rankings or raw times of athletes, pertaining to each individual component, may not reveal a great deal of information as to which individual discipline plays the most important role in deciding whether a triathlete can win the competition. This is due to the immediate succession of the events: what absolute times or rankings are achieved by the triathletes in each component are not vital. It is the time difference behind the lead athlete in each component that matters. Ranking second by a large time difference in any triathlon component implies a rather small chance for winning the contest compared to ranking fourth or fifth in any discipline with a much smaller time difference. Therefore, preprocessing the times and converting them into differential times becomes crucial to data mining in the triathlon. This is the approach taken by Ofoghi, Zeleznikow, and MacMahon (2011b) for analyzing triathlon data. They found that with approximately 87% certainty, a male triathlete can only win a medal if their differential running time is $\leq$26 seconds. For female triathletes, they found the certainty level is approximately 86% and the affordable differential running time for winning a medal is 28 sec.

The effect of the number of players/athletes on descriptive conversion. When considering the effect of this feature on descriptive conversion in data preprocessing, one should take into consideration the interactions and collaborations that occur in single-player as compared to team sports.

Single-player sports (e.g., fencing or singles table tennis) may require data, in the form of the RTS measures, to be converted only to the extent that enables modeling the dynamics of the interactions between the current athlete against his or her opponent. In most of these cases, the ultimate goal of performance analysis is to discover the patterns that may directly lead to defeating the opponents. For instance, the analysis of singles tennis performance data is carried out to find the best strategies that will result in winning the game against the opponent. Therefore, there is little need for converting tennis data or extracting new types of explicit data in single-player sports.

Team sports, on the other hand, may necessitate data preprocessing toward modeling the collaboration that occurs among different players in a team or selecting variables that indicate team versus individual performance. In most such sports, while the ultimate goal of performance analysis is to provide information that can be used to enhance the chance of winning, there are also intermediate goals or moderators in terms of finding the best collaborative actions/reactions that may lead to more successful scenarios in certain places or times of a game. Considering the game as a whole, this will increase the winning chances of the team. Finding the best collaborative strategy to enhance moving the ball toward the dangerous 16-m penalty area in football is an example of such an analysis. In this case, data related to each athlete/player pertaining to different times and/or places in the game require specific treatment before any actual analysis. Depending on the particular problem, the status of the original data collected, and the capabilities of the data analytical method to be employed, one may need: (a) to assign predefined event labels to each data record (for each player) in the data collection (e.g., shots on goal in football data) and (b) to group all events that occur in succession or recursion (by looking at the timestamps) for further pattern mining.

In addition, when analyzing team sports, in contrast to single-athlete sports, data records may be sampled for different individual team members (e.g., defender vs. attacker or play-maker vs. striker or goalkeeper in hockey) or groups of members (e.g., attackers, defenders, or midfielders). Depending on the problem under study, these data may also require certain descriptive conversion tasks. For instance, to understand the common mistakes made by the defense of a football team, a coach may be interested in knowing the formation or position of midfielders and attackers at the times that certain mistakes related to them were made and find the best team strategy to reduce the number of times that the team concedes a goal. In this particular case, there will be a need for relating data records gathered for the players with respect to time of the contest. Although relating different data may not seem to be a direct descriptive conversion, it will eventually alter the primitive data records into more descriptive and complete data records on which more informative data analysis can be carried out.

The effect of the duration of an event on descriptive conversion. The duration of competitions is a factor that needs to be considered in combination with other specifics of sports when deciding what type or amount of data preprocessing is necessary. RTS measures gathered in fixed-time and fixed-distance single-component single-player sports (e.g., judo, karate, fencing, and running) can potentially be the main data necessary for performance analysis. This excludes multiple-component sports such as triathlons where, as mentioned before, times need to be converted to differential times. Other types of fixed-time and fixed-distance events like football, hockey, rugby, and rowing may require sophisticated preprocessing tasks, such as extraction, structural conversion, and descriptive conversion.

In most variable-time events, RTS measures (i.e., pure times and scores referenced to timestamps) are the main required data for actual data analysis. In tennis, for instance, it is not necessarily required to convert time data related to how long balls are used (in tournaments played under the International Tennis Federation rules, balls are changed after the first nine games, then after the next eleven games, then after the next nine games, then after the next eleven games, and so on.) or cumulative times of previous games into other formats. Similarly, the scores (rather than times) are self-indicative of performances of athletes or teams in most variable-time events. However the remaining time available in a football, hockey, or water polo game is very important.

In terms of multiple-effort sports, such as gymnastics or diving, RTS measures, as well as other performance measures (e.g., body conformance in still rings or the amount of water splash in diving, if explicitly available), are adequate evidence of athletes' performances and may not require specific conversions or the extraction of new explicitly available data.

The effect of winning criteria on descriptive conversion. Winning criteria in sports demonstrate their effect in terms of data preprocessing, especially descriptive conversion, in the way that data are gathered, maintained, and reported during and after major events.

In sports where scores decide the winner (e.g., karate and judo), in most cases, only the scores are reported. In some cases, such as hockey, scores are tagged with timestamps. Other statistics of games (e.g., forced/unforced turn-overs and lost/gained positions) are also reported. In general, however, if time-dependent aspects of the games are under question, then certain data need to be converted or extracted. In football, for example, if one wants to understand the likelihood of winning the game given that the team was behind in the first half, then analysts need to label other previous games with respect to the results of the first halves, a task that may not be part of the original data collection.

In sports where time is the winning criterion, such as marathon and middle-distance running, time measures corresponding to each athlete along with their personal characteristics and some physiological data are mainly measured and reported. In some cases, this requires an extensive search for athlete-specific data that are not explicitly reported. These might include birth places and birth dates of subject athletes. Preprocessing is necessary to analyze the more sophisticated aspects of these sports in terms of velocity and/or time-related distance. An example of such research is the work by Jones and Whipp (2002) that calculated all the time-referenced velocity and distances from the paths taken by the runners under study. They used slow motion video playback for gold and silver medalists at the 800-m and 5000-m races in the Sydney Olympic Games to find the total distances they covered during the races. They then used the official race times to calculate the average velocity the runners sustained over the distances they covered in the races.

The nature of the pre-set number of scores needed to be achieved in certain variable-time sports, such as in volleyball and tennis, makes the final scores required to win less important when compared to other types of sports. In this case, the emphasis is on during-the-game measures, such as the data related to individual athletes' performances, individual scenarios of the games, or the tactics that may lead to success. Some of these data are partially measured and reported during major competitions. Some data can only be extracted from existing and reported data. In terms of volleyball, the analysis of the probability of winning the game (given winning in certain sets, e.g., the first two sets) requires game tagging with respect to the performances of the teams in the first and second sets as well as their final result. The analysis of the capabilities of certain male tennis players to win in grand-slam tournament tennis games, when they are behind by two sets, for example, also requires preprocessing of existing data in terms of the results of each set in their games. This can be carried out by labelling their games with the number of sets they have been behind (0, 1, or 2) as well as their final result (won, lost). The individual scores in each set do not matter, only who won the set.

In multiple-effort sports, where only the scores of each attempt and the final scores are reported, any data treatment may depend on the specific problem under study. Similarly to the case of sports with pre-set numbers of scores (e.g., tennis and volleyball), multiple-effort sports (e.g., diving), may require the extraction of during-the-game measures from existing data before data analysis can be conducted. The analysis of the likelihood of finishing an event in first place given that the athlete had not performed their best in their first attempt is an example of such analysis that requires extra data tagging with respect to the first attempts and the final standings of the athletes.

Sports Data Analysis

As mentioned earlier, the main data mining methods that have been used in the sports domain are clustering, classification, relationship modeling, and rule mining. We believe that it is ultimately the questions posed in certain sports that determine the choice of data mining method to be employed for performance analysis. However, the specifics of each sport category or individual sport must also be taken into consideration.

From a sport science viewpoint, the analysis of sports performance data, in the form of match analysis, which is the analysis of important attributes, events, strategies, or patterns (and their importance) in gaining success in different contests, is carried out with one (or a combination) of the following aims:

- Finding performance patterns that describe how an athlete or a team may increase their chances of finishing a competition in a certain position (e.g., boats that win standard 2000-m rowing races mostly finish each of the first three 500-m sectors in the fastest time, but this is not necessarily true for the last sector; Ofoghi et al., 2011a).
- Predicting performances of an athlete or a team given information related to their prior performances or training sessions (e.g., a sport performance analyst may predict the Olympic competitive performance of a female swimmer given previous swimming performances; Edelmann-Nusser et al., 2002).
- Real-time decision-making on what actions/reactions or strategies are required in the course of a current event (e.g., how to adjust the positioning of football players on the field when the team is one player short and one goal behind in the last 10 min of a critical game or what effort is required to gain a medal after the elimination race in the six event cycling omnium; Ofoghi, Zeleznikow, MacMahon, & Dwyer, 2011c).
- Finding the main demands of certain sport competitions and selecting athletes who can best address the demands (e.g., the track cycling omnium is better performed by riders with higher expertise in sprint-based events such as the flying time trial; Ofoghi et al., 2010).

While the first three aims are mostly related to short-term strategic planning for achieving success in forthcoming or current events, the fourth aim is mostly concerned with a long-term process towards talent identification, talent transfer (from one sport to another), and athlete development to secure success in future competitions.

From a data analytical viewpoint, each method within data mining (i.e., clustering and classification) can be implemented using different techniques (i.e., k-means, decision trees, Bayesian networks, support vector machines, and regression analysis). Each technique can be characterized in terms of three major characteristics:

1. Interpretability: how easily the results achieved by employing a certain method can be interpreted by data analytics experts and understood by (sport) professionals who are not experts in the data analysis domain.
2. Precision: the accuracy and reliability of the results that are derived using this technique.

3. Flexibility: the degree to which a certain method can be utilized for analyzing certain problems with different parameters and/or different data.

Each of the aforementioned goals in sports performance analysis necessitates the use of an appropriate specific data analytical method while each data mining technique for that specific method has its own characteristics in terms of interpretability, precision, and flexibility. On the other hand, each sport performance analysis requirement demands different levels from each technique. To better describe these points, we consider a rectangular model, as shown in Figure 1b, for sports performance analysis.

In this model, it is necessary to define two mappings in order to carry out insightful performance analysis tasks: (a) the mapping between the data mining methods and the sports performance analysis requirements and (b) the mapping between the sports performance analysis requirements and the data mining technique characteristics. The mapping between the data mining methods and data mining techniques, and also, the mapping between the data mining techniques and the data mining technique characteristics, fall mostly in the computer science domain and, therefore, are beyond the scope of this study.

The left side of Table 2 shows the mapping that we draw between the data mining methods and the sports performance analysis requirements. Performance pattern discovery is a task that is mainly carried out using clustering techniques and validated by utilizing classification systems. Clustering techniques, in particular, are used when the underlying structure of performance has a major unknown component that is to be discovered. Examples of using clustering techniques for extracting performance patterns are discussed in literature (Chen, et al., 2007; Lamb, Bartlett, & Robins, 2010; Ofoghi et al., 2010; Woolf, Ansley, & Bidgood, 2007).

Performance prediction, in comparison, is a task that is mainly addressed using classification systems and relationship modeling. Classification systems are suitable because of their ability to predict an already known target class label for unseen/unknown data instances. In the sports performance analysis context, once a classification system is trained with labeled performance data, it is then possible to predict the class label for new (previously unseen) data records for which

TABLE 2
The Mapping between Sports Performance Analysis Requirements, Data Mining Methods and Data Mining Technique Characteristics

	DM Methods				*DM Technique Characteristics*		
Sports performance analysis requirements	*Clust.*	*Class.*	*Rel. Mod.*	*Rule M.*	*Interp.*	*Prec.*	*Flex.*
Performance pattern discovery	✓	✓	—	—	high	moderate	moderate
Performance prediction	—	✓	✓	—	low	high	high
Real-time decision-making	—	—	✓	✓	very high	high	very low
Demand analysis	✓	✓	—	—	moderate	moderate	Moderate

Note. DM = data mining; Clust. = clustering; Class. = classification; Rel. Mod. = relationship modeling; Rule M. = rule mining; Interp. = interpretability; Prec. = precision; Flex. = flexibility.

A checkmark indicates that the specific performance analysis is mainly performed using the method in that column. A dash shows that the specific method in that column is not frequently and effectively used for the specific performance analysis in that row.

there is no target performance class assigned—an instance of unsupervised learning. Predicting the performance level of rowers into three known categories novice, good (sub-elite), and elite conducted by Smith and Spinks (1995) using linear discriminant analysis is a good example of such an analysis.

Another avenue to performance prediction is relationship modeling between predictor attributes and the dependent variable that represents performance. Major examples of this approach include the studies carried out by Edelmann-Nusser et al. (2002); Johnson, Edmonds, Jain, and Cavados (2009); Kahn (2003); Shao (2009), and Wilson et al. (2001).

Real-time decision-making (e.g., changing the line-up or field positioning during a game based on the progress during that contest) should be addressed using relationship modeling methods of data mining. Although there are sports in which some basic descriptive statistics are carried out and immediate inferences are made (e.g., having three inside 50s in a whole half of Australian rules football indicates poor performance, a low score, and very little chance of winning, thus a key focus for the second half), it is difficult to perform deeper analyses in real time.

Clustering, classification, and even rule mining methods tend to produce results that can better be used prior to major events, such as winning patterns, success event association and sequential patterns, and certain performance likelihoods. Relationship modeling, in contrast, can be employed in real time to integrate existing evidence based on the conditions and specifics of the current event and produce likelihoods of certain outcomes based on which to readjust (e.g., finding the best strategy to maintain the lead in a football match in the last 20 min given the opponent is not aggressively attacking).

There is a close relationship between performance pattern discovery and demand analysis. Demand analysis can also be more effectively conducted using clustering and classification techniques. The relationship between the two tasks is mainly due to the fact that performance patterns are, in many cases, among the most evident pieces of information that impose demands on athletes to participate in specific sport competitions. For instance, if in a triathlon, the winning pattern implies finishing the running component with the best performance, then the main demand of this sport (i.e., the key contributor to success in this sport) is to be a strong runner. Demand analysis for certain sports or competitions may also be carried out in terms of other pieces of information such as required prior training, nutrition, and physical strength; demands that are not necessarily revealed in performance patterns alone.

The right side of Table 2 shows our mapping between the sports performance analysis requirements and the data mining technique characteristics. In developing this mapping, we considered three main aspects: (a) the amount of output information generated by the analysis, (b) the rate of the reliance on the results by coaches and athletes which is defined as how core the results are to making important decisions, and (c) the time-frame within which the results produced by the analysis are to be utilized. These three aspects influence the extent to which the characteristics of the three techniques are required for sports performance analysis.

Processes that generate large amounts of output information generally need higher levels of interpretability of the results. This enables end users (i.e., those who may have less computer science expertise, such as sports coaches) to better understand and make use of the large amounts of results of the analysis (e.g., the average RTS measures required for finishing in second position in rowing). Whereas in cases where the output information is only a predicted class label, there is less need for perfect interpretability in the results. As an example, in performance prediction in Alpine skiing, the output information can be limited to predicted performances, in terms of medal winner or non-medal winner rankings, which do not necessitate much interpretability.

The rate of the reliance on the results mainly indicates the level of precision that a data mining analysis requires. Results that are core to decision making and are thus anticipated to be relied on very heavily require a high level of precision. In some analyses (e.g., performance pattern discovery), although the highest accuracy is desired and the output information is insightful, there is an opportunity for employing and developing alternative plans that reduces the rate of reliability of the results and therefore the required level of precision.

The time-frame within which to utilize the results of the analyses directly affects the required level of flexibility characteristic of a technique. The longer the time-frame for utilizing the results of the analyses, the more flexibility is desired. Therefore, it is useful to conduct a series of experiments with different settings and data, testing and refining the results. In short-term processes (e.g., real-time decision-making tasks), less flexibility will damage the effectiveness and efficiency of the performance analysis task to a lesser extent.

CONCLUSION

This article provides a narrative review of the literature on data mining and sports, and a framework for the next steps in bridging the two domains. Each performance analysis task in elite sports requires different data preprocessing and data analysis techniques. Data preprocessing is most influenced by the category and feature of sports (e.g., the number of individual events in a competition, the number of players in the games, the duration of the games, and the winning criteria). These necessitate different preprocessing tasks.

Data analysis that comes after data preprocessing is more influenced by the type of problem being tackled in the sports performance analysis. Performance pattern discovery, performance prediction, real-time decision-making, and demand analysis problems are often better carried out using different data mining methods and require different interpretability, precision, and flexibility measures.

To cover all of the aspects of sports performance analysis in a general structure, we presented a rectangular model bringing together performance analysis requirements, data mining methods, data mining techniques, and technique characteristics. This inter-connected rectangular model requires sufficient attention before conducting practical and useful performance analysis tasks. The mappings that we discussed between some of the main elements in this model suggest what data mining methods and techniques are suitable for which sports performance analysis problems.

Future research should empirically validate the model as well as provide evidence that data mining, in comparison with alternative tools, can inform athletes and coaches to perform better. Further, the adaptation and development of data mining tools that fit the specific needs of the sport domain are warranted allowing fast and accurate estimates of previous, current and future behavior. These activities require networks of research and practice allowing sport and data mining knowledge to be combined respecting ethical rules, long-standing archives and communication between the stakeholders.

Our review on the different data analytical demands of different elite sports is an unprecedented effort to shed more light on different aspects of the use of sophisticated data analysis and mining methods in the domain of sports performance analysis. This will assist both data analysts and sport professionals to more effectively collaborate and enhance their understanding of a variety of participant factors that contribute to success in sport events at different levels.

ACKNOWLEDGEMENTS

We thank the Institute of Sport, Exercise, and Active Living at Victoria University, Melbourne, Australia, for funding this research project. We also thank a number of sport scientists at the Victorian Institute of Sport, Australia, for their great support.

REFERENCES

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In P. S. Yu & A. L. P. Chen (Eds.) *Proceedings of the Eleventh International Conference on Data Engineering* (pp. 3–14). Taipei, Taiwan: IEEE Computer Society.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, *22*, 207–216. doi: 10.1145/170036.170072

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed). Cambridge, MA: MIT Press.

Anagnostopoulos, T., Anagnostopoulos, C. B., Hadjiefthymiades, S., Kalousis, A., & Kyriakakos, M. (2007). Path prediction through data mining. In *IEEE International Conference on Pervasive Services* (pp. 128–135). Istanbul, Turkey: IEEE.

Berkhin, P. (2006). A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, & M. Teboulle, *Grouping multidimensional data* (pp. 25–71). Berlin, Germany: Springer.

Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., & Ramanujam, K. (1997). Advanced Scout: Data mining and knowledge discovery in NBA data. *Data Mining and Knowledge Discovery*, *1*, 121–125. doi: 10.1023/A:1009782106822

Bishop, D. (2003). Performance analysis: What is performance analysis, and how can it be integrated within the coaching process to benefit performance? *Peak Performance*, 4–7. Retrieved December 2010, from http://www.pponline.co.uk/encyc/sports-performance-analysis-.

Chen, I., Homma, H., Jin, C., & Yan, H. H. (2007). Identification of elite swimmers' race patterns using cluster analysis. *International Journal of Sports Science and Coaching*, *2*, 293–303. doi: 10.1260/174795407782233083

Clausen, H. (2012). *Important concepts of machine learning* (1st ed.). New Delhi, India: World Technologies.

Cox, T. F., & Dunn, R. T. (2002). An analysis of decathlon data. *Journal of the Royal Statistical Society*, *51*, 179–187. doi: 10.1111/1467-9884.00310

Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modelling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science*, *2*, 1–10. doi: 10.1080/17461390200072201

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. doi: 10.1145/240455.240464

Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis*. Burlington, MA: Academic.

Johnson, M. B., Edmonds, W. A., Jain, S., & Cavados, J., Jr. (2009). Analysis of elite swimming performances and their respective between-gender differences over time. *Journal of Quantitative Analysis in Sports*, *5*(4). doi: 10.2202/1559-0410.1186

Jones, A. M., & Whipp, B. J. (2002). Bioenergetic constraints on tactical decision making in middle distance running. *British Journal of Sports Medicine*, *32*, 102–104. doi: 10.1136/bjsm.36.2.102

Kahn, J. (2003). *Neural network prediction of NFL games*. Retrieved from http://homepages.caewisc.edu/~ece539/project/f03/kahn.pdf?q-nf1-search-statistics.

Kenny, I. C., Sprevak, D., Sharp, C., & Boreham, C. (2005). Determinants of success in the Olympic decathlon: Some statistical evidence. *Journal of Quantitative Analysis in Sports*, *1*(1), article 3.

Kline, C. E., Durstine, J. L., Davis, J. M., Moore, T. A., Devlin, T. M., Zielinski, M. R., & Youngstedt, S. D. (2007). Circadian variation in swim performance. *Journal of Applied physiology*, *102*, 641–649. doi: 10.1152/japplphysiol.00910.2006.

Lamb, P., Bartlett, R., & Robins, A. (2010). Self-organising maps: An objective method for clustering complex human movement. *International Journal of Computer Science in Sport*, *9*, 20–29.

Liao, T. (2008). Tactics analysis on women swimming athletes in the 800m freestyle swimming race using speed coefficient theory. In Q. Lo & B. K. M. Sim (Eds.) *Proceedings of International Workshop on Knowledge Discovery and Data Mining* (pp. 453–456). Washington, DC: IEEE Compter Society. doi: 10.1109/WKDD.2008.145

Ofoghi, B., Stefano, D., Zeleznikow, J., & MacMahon, C. (2012). Modelling relationships between swimming attributes for performance prediction. In *American College of Sports Medicine*. San Francisco, California USA, May 29–June 2.

Ofoghi, B., Zeleznikow, J., & MacMahon, C. (2011a). Probabilistic modelling to give advice about rowing split measures to support strategy and pacing in race planning. *International Journal of Performance Analysis in Sport*, *11*, 239–253.

Ofoghi, B., Zeleznikow, J., & MacMahon, C. (2011b). A machine learning approach to triathlon component analysis. In Y. Jiang & H. Zhang (Eds.) *Proceedings of the International Symposium on Computer Science in Sport* (pp. 30–33). Shanghai, China.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Dwyer, D. (2010). A machine learning approach to predicting winning patterns in track cycling omnium. In M. Bramer (Ed.) *Proceedings of the International Federation for Information Processing (IFIP) Conference on Advances in Information and Communication Technology* (pp. 67–76). Brisbane, Australia: Springer Berlin Heidelberg.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Dwyer, D. (2011c). Has the addition of the elimination race to the track cycling omnium benefitted sprinters or endurance riders?. In Y. Jiang & H. Zhang (Eds.) *Proceedings of the International Symposium on Computer Science in Sport* (pp.34–37). Washington, DC: IEEE Computer Society.

Pollard, G., & Pollard, G. (2010). Four ball best ball 1. *Journal of Sports Science and Medicine*, *9*, 86–91.

Ransdell, L. B., Vener, J., & Huberty, J. (2009). Masters athletes: An analysis of running, swimming and cycling performance by age and gender. *Journal of Exercise Science and Fitness*, *7*, S61–S73.doi: 10.1016/S1728-869X(09)60024-1

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, *33*, 163–180. doi: 10.1177/0165551506070706

Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). *Sports data mining*. *Integrated series in information systems* (Vol. 26). New York, NY: Springer.

Shao, S. (2009). Application of BP neural network model in sports aerobics performance evaluation. In F. Xiong & Q. Li (Eds.) *Proceedings of the 2009 Pacific-Asia Conference on Knowledge Engineering and Software Engineering* (pp. 33–35). Washington, DC: IEEE Computer Society.

Smith, R. M., & Spinks, W. L. (1995). Discriminant analysis of biomechanical differences between novice, good and elite rowers. *Journal of Sports Sciences*, *13*, 377–385.

Smyth G.K. (2002). Nonlinear regression. *Encyclopedia of Environmetrics*, *3*, 1405–1411.

Stranieri, A. & Zeleznikow, J. (2005). *Knowledge discovery from legal databases: Law and Philosophy Library*(Vol. 69). Dordrecht, The Netherlands: Springer.

Sun, J., Yu, W., & Zhao, H. (2010). Study of association rule mining on technical action of ball games. In *Proceedings of the 2010 International Conference on Measuring Technology and Mechatronics Automation (ICMTMA 2010*; pp. 539–542). Washington, DC: IEEE Computer Society.

Vaz, L., Rooyen, M. V., & Sampaio, J. (2010). Rugby game-related statistics that discriminate between winning and losing teams in IRB and super twelve close games. *Journal of Sports Science and Medicine*, *9*, 51–55.

Vezos, N., Gourgoulis, V., Aggeloussis, N., Kasimatis, P., Christoforidis, C., & Mavromatis, G. (2007). Underwater stroke kinematics during breathing and breath-holding front crawl swimming. *Journal of Sport Science and Medicine*, *6*, 58–62.

Wilson, B., Mason, B., Cossor, J., Arellano, R., Chatard, J., & Riewald, S. (2001). Relationships between stroke efficiency measures and freestyle swimming performance: An analysis of freestyle swimming events at the Sydney 2000 Olympics. In J. R. Blackwell & R. H. Sanders (Eds.) *Proceedings of Swim Sessions*: *XIX International Symposium on Biomechanics in Sports* (pp. 79–82). San Francisco. CA: University of San Francisco.

Woolf, A., Ansley, L., & Bidgood, P. (2007). Grouping of decathlon disciplines. *Journal of Quantitative Analysis in Sports*, *3*(4), article 5. doi: 10.2202/1559-0410.1057

Zwols, Y., & Sierksma, G. (2009). Training optimization for the decathlon. *Journal of Operations Research*, *57*, 812–822. doi: 10.1287/opre.1080.0616