Data Analysis and Mining Demands in Elite Sports

Journal:	*Measurement in Physical Education and Exercise Science*
Manuscript ID:	Draft
Manuscript Type:	Original Article
Keywords:	Statistics, Evaluation, exploratory factor analysis < Factor Analysis

Running head: DATA MINING DEMANDS IN SPORTS

Data Analysis and Mining Demands in Elite Sports

—

—

Abstract

Sophisticated data analytical methods such as data mining, with an exploratory focus, are becoming increasingly useful tools in analyzing sports performance data and supporting decision making that is crucial to gaining success in elite sports. In this article, we investigate the different data mining demands of different elite sports with respect to a number of features that describe sport competitions. Our main aim is to more formally connect the sports and data mining domains through: (a) describing a framework for categorizing sports with respect to pre-specified features, and (b) better understanding the analytical demands of different types of performance analysis. For this, we review different aspects such as sport categories and performance analysis requirements, that influence certain stages in sports data mining.

Keywords: Data mining, sport, performance analysis

Data Analysis and Mining Demands in Elite Sports

Performance analysis, as a means for creating and analyzing a valid record of athlete performances using systematic observations, has gained importance in the last decade. It has been facilitated by advances in information technology and digital photography (Bishop, 2003). This type of analysis can be either in the form of *match analysis* or *biomechanics*. Match analysis is more concerned with analyzing important attributes, events, strategies, or patterns (and their importance) in gaining success in different contests whereas biomechanics is more focused around the application of mechanical principles to human's biological system and how any improvements in (any element of) this system may result in desired success in sports. Performance analysis, in either form, can significantly assist coaches and athletes at any level in the process of decision-making during or prior to sport competitions.

Despite the complexity and obvious demands of sport decisions, much of the coaches' decision-making in sport has been based on tradition and emulation (Williams & Ericsson, 2005). Traditional coaching may, for instance, dictate that large amounts of immediate feedback on learning and skills practiced in blocked fashion will lead to improvements. However, research controlling the structure of practice and feedback indicates otherwise (Magill, 2007). Although, in most cases, there are data pertaining to performance and training, systematic analysis and sophisticated data analytical approaches, which may help avoid common errors such as base-rate neglect (Fiedler, Brinkmann, Betsch, & Wild, 2000), have not been applied to gaining greater insight and support for these tasks in an influential way.

We believe the main challenges that to date interfere with large-scale and influential utilization of effective analytical approaches in the sports performance analysis include: (a) the lack of a belief that the outcomes of performance analysis can practically improve future athlete performances in major competitions, and (b) the lack of a well-defined data analytical framework that takes into consideration the main demands of different problems

in elite sports and sport performance analysis requirements, and (c) the lack of confidence and mutual understanding between sports data analysts and professional coaches/athletes, especially in cases where the two come from different knowledge backgrounds.

While the first and third items rely on the cultural and psychological aspects of integrating sports and data analytical procedures, the second item relates to guidance for appropriate methods. A domain of methods for analyzing sports data is data mining which is a branch of the broader domains of computer science and artificial intelligence.

Data mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Using data mining techniques, useful and previously unknown information can be extracted from archived or streaming data. The extracted information may be in the form of prediction of future events, finding events that co-occur and their sequence of occurrence, and grouping entities that are similar according to known attributes.

Unlike many other applications of data analysis and mining (e.g., customer analysis, agricultural analysis, surveillance, etc.), elite sports data analysis must be matched, case by case, with the demands of the given sport or sport task. This makes the domain of sports data analysis a more sophisticated area of application (than aforementioned problems) for analytical methods where converted measures and numbers require special treatments regarding certain demands.

Regardless of physiological studies of recovery methods, biomechanical studies of techniques, or motor learning studies of training interventions, this article focuses on the analysis of archived performance data using data mining (i.e., pre-processing and data analysis) techniques and therefore tackles the second item above with two main aims: (a) to describe a framework for categorizing different sports with respect to pre-specified attributes that relate to performance analysis, and (b) to better understand the analytical demands of different types of research problems in elite sports in terms of performance

analysis.

In keeping with the characteristics of data mining, the major goal of this study is to focus on the *exploratory* level of data analysis. Thus, the extraction of useful information from previous data is our major concern. The article, therefore, leaves aside the *explanatory* level that explains and justifies any useful information discovered from the data.

Data Mining in Elite Sports

Most of the performance analysis work in sports, concerning match analysis, has been based on using statistical measures. These approaches range from straightforward to sophisticated analyzes. The 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., calculating correlation measures between the values of two variables without considering other conditions that may have lead to certain values of the two variables). Examples of studies that are more in the straightforward data analysis category include the works by Pollard and Pollard (2010); Ransdell, Vener, and Huberty (2009); Vezos et al. (2007).

In sophisticated approaches, however, more in-depth analysis is carried out to find 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 the studies by Cox and Dunn (2002); Kenny, Sprevak, Sharp, and Boreham (2005); Kline et al. (2007); Liao (2008); Vaz, Rooyen, and Sampaio (2010); Zwols and Sierksma (2009).

Although the above-mentioned studies have shed light on different aspects of a number of elite sports, the use of statistical analysis without an understanding of their fundamental meaning (and the contextual backgrounds of variables/values) under study in terms of certain sports can potentially be misleading due to impreciseness or over-emphasis

of the statistics (Schumaker, Solieman, & Chen, 2010). Impreciseness may be the result of missing/ignoring other contributing variables and over-emphasis can be due to ignoring the effect of conceptually non-related coincidences. The utilization of data mining techniques is a way to overcome these problems.

In the sports domain, data mining methods have generally been used to model the inter-relationships of performance measures and attributes and to also extract athlete performance patterns from previously held competitions. These can be used in the decision-making processes to support *strategic planning* and *athlete selection*.

Major Data Mining Methods Used in Sports Domain

Clustering. Clustering, also referred to as unsupervised learning in the machine learning literature, is concerned with finding the underlying structure in a set of data points where there is no (prior) label information available for the data points. The result of cluster analysis is 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. Unsupervised learning may be used for data reduction, pattern discovery, and outlier detection. Table 1 summarizes the other major uses of clustering for sport data analysis.

Classification. Classification, also known as supervised learning in the machine learning literature, is used to predict group memberships for data instances (i.e., individual cases) in a data set. The output of classification is a set of data instances that are labeled with pre-defined and existing class labels. This task relies on known structures or knowledge that can then be applied to unseen new data. Table 1 summarizes some of the other previous studies that made use of classification techniques for sports performance analysis.

Relationship modeling. The aim of relationship modeling in a complex environment with a number of participant attributes is to find a function or model that can define the

data and describe the inter-relationships 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 is one of the methods that has been used for fitting a line or polynomial to data. A limitation of regression analysis is that, in most cases, it is not appropriate for modeling data where there is a large volume of non-linearity involved in the underlying relationships. To overcome this, some researchers have made use of non-linear methods, such as Neural Networks, to find attribute inter-relationships. Some of the previous studies that utilized 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 and being physically weak) has two main forms in data mining, namely association rule mining (Agrawal, Imielinski, & Swami, 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, 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 in a broad sense. Association rules and sequential patterns have been applied in a few studies (Bhandari et al., 1997; Jing, Wenshuang, & Huiqun, 2010).

Current Status of Sports Data Mining

In general, the utilization of different data mining methods (i.e., clustering, classification, relationship modeling, and rule mining) in the sports domain suggests that sports data mining can now be recognized as a distinct area of knowledge that requires more in-depth attention by researchers in both domains of sports science and computer

science. However, to date, there has not been significant effort from either side to more formally and structurally connect the two domains. 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 in limited contexts. Making the connection between the two domains of knowledge (i.e., sport science and computer science) will help avoid common errors (e.g., see (Fiedler et al., 2000)) and more effectively and efficiently handle sports performance data.

In the next sections, we will go beyond this boundary and provide a more general structure that introduces sports categories and types of research problems that are encountered in the sports performance analysis domain, which we believe will facilitate future research in this domain.

Categorization of Sports

There are a number of methods of classifying sports using dimensions such as the environmental context (i.e., open versus closed) and primary actions (e.g., striking versus endurance). Our goal here is to highlight the most relevant features from a data mining point of view. Before any data analytical approaches can be utilized for performance analysis in sports, it is necessary to understand the specifics of the certain sport under study, and the most relevant questions that arise. For conducting data analysis and mining tasks on sport performances, we categorize sport competitions regarding a few sports-related aspects including the number of events, number of players/athletes, duration of the competition, and winning/evaluation criteria. The main motivation behind using these features is that they imply different types and methods of data mining especially in the data pre-processing step. This will be discussed later in this paper.

Sport-Related Data Mining Demands

When analyzing sports performance data, there are three main attributes that are of interest to sport scientists and professionals: (a) rankings, (b) times, and (c) scores. This is

mainly due to the fact that, apart from referees' discretion mentioned earlier, these three measures represent performance and are used for evaluating athletes in most sport events. We refer to these performance measures as RTS (ranking, time, score) measures.

We believe that sports performance analysis implies a mapping, as shown in Figure 1(a), from the sports domain to the data analytical domain (the data mining domain in this work). 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 pre-processing 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 due to the lack of precision, sound approach, or validity.

The mapping that occurs from the sports domain to the data mining domain for performance analysis opens a gap between the two domains of science, namely sport science and computer science. To minimize the effect of this gap, the mapping has to enable the possibility of inheriting 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 them.

Sports Data Pre-processing

Sports-related data pre-processing, in the data mining space, involves preparing and sorting data for analysis and can take one or some of the following forms, depending on the format of the data and the research problem:

(a) **Filtering:** Filtering data records considering certain categories of competitions e.g., categorizing rowing performance results in to fast, medium, and slow courses.

(b) **Format conversion:** Converting data into a format which can be interpreted by

specific data analytical software and tools used when conducting actual data analysis e.g., converting hh:mm:ss times into collective seconds.

(c) **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-meter rowing races based on the times of the boats (Ofoghi, Zeleznikow, & MacMahon, 2011b).

(d) **Structural conversion:** Converting parts of data into a format that allows for more precise data analytical process e.g., generalizing final standings ranging from 1 to the number of contestants into medal winner (positions 1 to 3) and non-medal winner (positions greater than 3) categories.

(e) **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 lead and other athletes.

The data pre-processing tasks in *filtering*, *format conversion*, and *extraction* are not affected by the sport specifics, while *structural conversion* is usually dependent on: (a) the specifics of the data analytical method where some data analytical methods can more effectively handle nominal values compared to numeric data types e.g., classification systems, and (b) the amount of data that is available where often small amounts of data pertaining to specific parameters may necessitate generalization of the values into more coarse-grained values that cover a greater number of data records. *Descriptive conversion* is, on the other hand, closely linked to understanding the sport and its features. As mentioned earlier, the magnitude of the effect that the sports domain may have on this type of pre-processing depends on the specifics of the sports and the sports categories. The appropriate data pre-processing technique may be chosen or driven by the combination of features of a sport e.g., scoring system, duration, number of events, etc. However, 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 pre-processing tasks may be appropriate.

The effect of the number of events in a competition on sports data pre-processing. Most single-component sports (e.g., fencing and running) usually do not require or influence descriptive conversion in data pre-processing. The main reason for this is, in these sports, the predictor variables i.e., the RTS measures, explicitly represent the nature of the sport and 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-meter freestyle swimming does not require any specific type of pre-processing of lap times to predict finishing times.

Multiple-component competitions, 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). The three performance measures, RTS, in independent multiple events usually provide enough information for modeling and predicting dependent variables i.e., the overall times and the overall final standings. The machine learning-based analysis of the necessary abilities for winning the track cycling omnium competitions carried out by Ofoghi, Zeleznikow, MacMahon, and Dwyer (2010), for example, includes no further pre-processing on the rankings of riders in each individual event than generalizing the final standings into the three categories of medal winners, non-medal winners ranked between 4 and 10, and non-medal winners ranked above 10. This is mainly due to the small amount of historical data which is available for this specific sport.

In successive multiple-event sports, in contrast, raw RTS measures may be misleading in the performance analysis task. 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 where it does not matter what absolute times or rankings are achieved by the triathletes, but the time difference behind the lead athlete in each component. 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, pre-processing the times and converting them into differential times becomes crucial to data mining in the triathlon. This is the approach Ofoghi, Zeleznikow, and MacMahon (2011a) have taken for analyzing triathlon data and found that with a ∼87% certainty, a male triathlete can only win a medal if their differential running time is ≤26 seconds. For female triathletes, they found the certainty level is ∼86% and the affordable differential running time for winning a medal is 28 seconds.

The effect of number of players/athletes on sports data pre-processing. The number of players/athletes cannot be used alone to decide what type of pre-processing needs to be conducted. However, when considering the effect of this feature on descriptive conversion in data pre-processing, one should take into consideration the different characteristics of the *interactions* and *collaborations* that occur in single-player versus team sports.

Single-player sports (e.g., fencing and 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/her opponent/s. 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 tennis performance data is carried out to find the best strategies that will result in winning the game against the opponent. Therefore, there is a 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 pre-processing towards 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 re/actions that may lead to more successful scenarios in certain places or times of a game. Taking the game as a whole, this will increase the winning chances of the team. Finding the best collaborative strategy to enhance moving the ball towards the dangerous 16-meter penalty area in football is an example of such an analysis. In these cases, data related to each athlete/player pertaining to different times and/or places in the game require specific treatment before any actual analysis. For this, 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 pre-defined event labels to each data record (for each player) in the data collection (e.g., score or non-score labels in hockey data or shot 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, different data records may be sampled for different individual team members (e.g., in hockey, defender vs. attacker, play-maker vs. striker and goalkeeper) or groups of members (attackers, defenders, midfielders, etc.). 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 that relate 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.

The effect of the duration of an event on sports data pre-processing. The duration of competitions is a factor that needs to be considered in combination with other specifics of the sports when deciding about the type or amount of necessary data pre-processing. The

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 events 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 pre-processing tasks such as those mentioned earlier in this paper.

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 balls (in tournaments played under the International Tennis Federation rules, balls are changed after the first 9 games, then after the next 11 games, then after the next 9 games, then after the next 11 games etc.) 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 such sports.

In terms of multiple-effort sports, such as gymnastics and diving, the 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 adequately evident of athletes' performances and may not require specific conversions or the extraction of new explicitly available data.

The effect of winning criteria on sports data pre-processing. Winning criteria in sports demonstrate their effect in terms of data pre-processing, 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, 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 were behind in the

first half, then analysts need to label other previous games with respect to the results of the first halves, a task which 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 (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. Pre-processing is necessary to analyze the more sophisticated aspects of these sports in terms of velocity and/or time-related distance. An example of this is the work by Jones and Whipp (2002) in which they calculated all the time-referenced velocity and distances from the paths taken by the runners under study.

The nature of the pre-set number of scores to be achieved in variable-time sports, such as in volleyball and tennis, makes the final scores to win less important compared to other types of sports. This then puts the emphasis 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 and some 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 in winning the grandslam tennis games when they are behind by two sets, for example, also requires pre-processing of existing data in terms of the results of each set in their games. This can be carried out by labeling their games with the number of sets they have been behind (0, 1, or 2) as well as their final result (won, lost).

In multiple-effort sports, where only the scores of each attempt as well as the final scores are reported, any data treatment may depend on the specific problem under study.

Like in sports with pre-set numbers of scores (e.g., tennis and volleyball), in multiple-effort events (e.g., diving), it may be necessary to extract during-the-game measures from existing data before the data analysis step. The analysis of the likelihood of finishing an event in 1^{st} 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 drive 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, is carried out with one (or a combination) of the following aims:

(a) 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-meter rowing races mostly finish each of the first three 500-meter sectors in the fastest time, but this is not necessarily true for the last sector (Ofoghi et al., 2011b).

(b) Predicting performances of an athlete or a team given information related to their prior performances or training sessions e.g., a sport performance analyst might discover that the rider whose horse has not participated in any major competition in the last three months does not have a large chance of finishing a major horse riding race in the top three positions.

(c) Real-time decision-making on what re/actions or strategies to take 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 minutes of a major game.

(d) 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 items are mostly related to *short-term* strategic planning for achieving success in forthcoming or current events, the fourth item 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 (classification, clustering, etc.) can be implemented using different techniques and each technique can be characterized in terms of three major characteristics:

i. Interpretability: how easily the results achieved by employing a certain method can be interpreted by data analytical experts and understood by (sport) professionals who are not experts in the data analysis domain.

ii. Precision: how accurate and reliable are the results that are derived using this technique.

iii. 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 leads to the necessity of using 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 of each technique characteristic. Therefore, to better describe this need, we consider a rectangular model, as shown in Figure 1(b), 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 out of the scope of this study.

Table 2 (the top part) 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 the works in (Chen, Homma, Jin, & Yan, 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 data records for which there is no target performance class assigned. 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 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 in (Edelmann-Nusser, Hohmann, & Henneberg, 2002; Johnson, Edmonds, Jain, & Jr., 2009; Kahn, 2003; Shao, 2009; 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) is a task that has been addressed to a much lesser extent than other performance analysis areas. We believe the difficulty involved in this task originates from the main barriers for using (semi) automated data analytical methods in sports decision-making mentioned earlier. However, if this task is to be tackled, then we believe it should mainly 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 3 inside 50s a whole half, so obviously this is something to focus on in the second half), to perform deeper analyzes in real time, there is not much space for conducting those that produce results that are beneficial in the long term. 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 performances 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 re-adjust strategies, e.g., finding the best strategy to maintain the lead in a football match in the last 20 minutes given the opponent is not aggressively attacking.

There is a close relationship between performance pattern discovery and demand analysis; therefore, 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 triathlon, the winning pattern implies finishing the running component with the best performance, then the main demand (i.e., the key contributor to success in) of this sport is to have excellence in running. 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 that are not necessarily formulated

in performance patterns.

Table 2 (the bottom part) 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 amount of the three technique characteristics 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, for instance, 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 does not necessitate much interpretability.

The rate of the reliance on the results mainly indicates the level of precision that a data mining analysis requires to exhibit. 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 analyzes e.g., performance pattern discovery, although still the highest accuracy is desired and the output information is insightful, there is some space for employing and developing alternative plans (risk management). This reduces the rate of the reliability of the results and therefore the required level of precision.

The time-frame within which to utilize the results of the analyzes directly affects the required level of flexibility characteristic of a technique. The longer the time-frame of

utilizing the results of the analyzes is, the more flexibility is desired and 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; however, less flexibility will damage the effectiveness and efficiency of the performance analysis task to a lesser extent.

Conclusion

Performance analysis in terms of different elite sports implies different data pre-processing and data analysis techniques. Data pre-processing is more influenced by the category of sports where different features of the sports (i.e., the number of individual events in a competition, the number of players in the games, the duration of the games, and the winning criteria) necessitate different pre-processing tasks.

Data analysis that comes after data pre-processing 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 suit for which sports performance analysis problems.

Our review on the different data analytical demands of different elite sports is an unprecedent 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 eventually 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.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. *SIGMOD Rec.*, *22*(2), 207–216, DOI: 10.1145/170036.170072. Available from http://doi.acm.org/10.1145/170036.170072

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the eleventh international conference on data engineering* (pp. 3–14). Taipei, Taiwan.

Ball, K., & Best, R. (2007). Different centre of pressure patterns within the golf stroke i: Cluster analysis. *Journal of Sports Sciences*, 757–770, DOI: 10.1080/02640410600874971.

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*(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. Available from http://www.pponline.co.uk/encyc/sports-performance-analysis-coaching-and-training-39

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*(3), 293–303, DOI: 10.1260/174795407782233083.

Cox, T., & Dunn, R. (2002). An analysis of decathlon data. *Journal of the Royal Statistical Society*, *51*(2), 179–187, DOI: 10.1111/1467-9884.00310.

Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science*, *2*(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.

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond baserate neglect and statistical format. *Journal of Experimental Psychology: General*, *129*(3), 399–418.

Gaudreau, P., & Blondin, J.-P. (2004). Different athletes cope differently during a sport competition: a cluster analysis of coping. *Personality and Individual Differences*, *36*(8), 1865–1877, DOI: 10.1016/j.paid.2003.08.017. Available from `http://www.sciencedirect.com/science/article/B6V9F-49TRJYD-9/2/3c8754c3b4426ec0e38d6032cd343bad`

Jaitner, T., Mendoza, L., & Schllhorn, W. (2001). Analysis of the long jump technique in the transitions from approach to takeoff based on time-continuous kinematic data. *European Journal of Sport Science*, *1*(5), 1–11.

Jing, S., Wenshuang, Y., & Huiqun, Z. (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). Changsha, China.

Johnson, M., Edmonds, W., Jain, S., & Jr., J. C. (2009). Analysis of elite swimming performances and their respective between-gender differences over time. *Journal of Quantitative Analysis in Sports*, *5*(4), Article 2, DOI: 10.2202/1559-0410.1186.

Jones, A., & Whipp, B. (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* (Tech. Rep.). University of Wisconsin Electrical and Computer Engineering Department.

Kenny, I., 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., Durstine, J., Davis, J., Moore, T., Devlin, T., Zielinski, M., & Youngstedt, S. (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 *Proceedings of international workshop on knowledge discovery and data mining* (pp. 453–456, DOI: 10.1109/WKDD.2008.145). Adelaide, Australia.

Magill, R. (2007). *Motor learning and control: Concepts and applications* (8th ed.). New York, NY: McGraw-Hill.

Ofoghi, B., Zeleznikow, J., & MacMahon, C. (2011a). A machine learning approach to triathlon component analysis. In *Proceedings of the international symposium on computer science in sport* (pp. 30–33). Shanghai, China.

Ofoghi, B., Zeleznikow, J., & MacMahon, C. (2011b). Probabilistic modeling to give advice about rowing split measures to support strategy and pacing in race planning. *International Journal of Performance Analysis in Sport*, *11*(2), 239–253.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Dwyer, D. (2010). A machine learning approach to predicting winning patterns in track cycling omnium. In *Proceedings of the international federation for information processing (ifip) conference on advances in information and communication technology* (pp. 67–76). Brisbane, Australia.

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

Ransdell, L., 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*(2), S61–S73, DOI: 10.1016/S1728-869X(09)60024-1.

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

Shao, S. (2009). Application of bp neural network model in sports aerobics performance evaluation. In *Proceedings of the 2009 pacific-asia conference on knowledge engineering and software engineering* (pp. 33–35). Shenzhen, China.

Smith, R. M., & Spinks, W. L. (1995). Discriminant analysis of biomechanical differences between novice, good and elite rowers. *Journal of Sports Sciences*, *13*(5), 377–385.

Vaz, L., Rooyen, M., & 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.

Watson, A. (1988). Discriminant analysis of the physiques of schoolboy rugby players, hurlers and non-team members. *Journal of Sports Sciences*, *6*(2), 131–140.

Williams, A. M., & Ericsson, K. A. (2005). Perceptual-cognitive expertise in sport: Some considerations when applying the expert performance approach. *Human Movement Science*, *24*(3), 283–307, DOI: 10.1016/j.humov.2005.06.002.

Wilson, B., Mason, B., Cossor, J., Arellano, R., Chatard, J.-C., & Riewald, S. (2001). Relationships between stroke efficiency measures and freestyle swimming performance: An analysis of freestyle swimming events at the sydney 2000 olympics. In *Proceedings of biomechanics symposia* (pp. 79–82). 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. Available from `http://ideas.repec.org/a/bpj/jqsprt/v3y2007i4n5.html`

Zwols, Y., & Sierksma, G. (2009). Training optimization for the decathlon. *Journal of Operations Research*, *57*(4), 812–822, DOI: 10.1287/opre.1080.0616.

Table 1
Utilization of clustering, classification, and relationship modeling techniques in the sports domain

Method	Researcher	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 organizing maps	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, Mendoza, & Schllhorn (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)

Table 2

The mapping between sports performance analysis requirements and major data mining methods and the mapping between the sports performance analysis requirements and the data mining technique characteristics

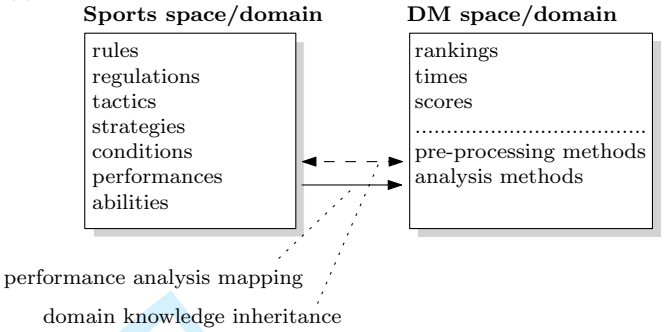
Sports performance analysis requirements	Data mining methods			
	clustering	classification	relationship modeling	rule mining
performance pattern discovery	√	√	–	–
performance prediction	–	√	√	–
real-time decision-making	–	–	√	√
demand analysis	√	√	–	–

	Data mining technique characteristics		
	interpretability	precision	flexibility
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

Figure Captions

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

(a)



(b)