Chapter 2
Sports Data Mining Methodology

Chapter Overview

Data Mining involves procedures for uncovering hidden trends and developing new data and information from data sources. These sources can include well-structured and defined databases, such as statistical compilations, or unstructured data in the form of multimedia sources such as video broadcasts and play-by-play narration.

Data mining activities, the tools, technologies and human expertise, are rooted within the field of Knowledge Management (Davenport and Prusak 1998). Knowledge Management can provide an organization with a means of competitive advantage (Lahti and Beyerlein 2000) and a method for maintaining the continuity of knowledge in the organization (Serenko and Bontis 2004). Through knowledge sharing and retention, businesses are discovering increased productivity and innovation (O’Reilly and Knight 2007). However, before getting raw data to a stage of usable knowledge, we must first examine the intermediate levels of data and knowledge as represented by the Data-Information-Knowledge-Wisdom (DIKW) hierarchy (Ackoff 1989). The DIKW hierarchy is a widely accepted concept in knowledge management circles as a way to represent different levels of what individuals see and know (Cleveland 1982; Zeleny 1987). Each successive level: data, information, knowledge and wisdom, builds upon prior levels and provides an increased awareness of surroundings (Carlisle 2006) where meaning can be found (Chen 2001, 2006).

The DIKW framework then sets the stage for disambiguating data from knowledge and sets definitional boundaries for what data, information and knowledge are. Applying this to the sports domain, certain activities and techniques operate at the data level (i.e., data collection, data mining and basic statistics). Other techniques and algorithms are more suited to the knowledge end of the spectrum, such as strategies and simulations. Throughout this chapter, the DIKW framework can be used to identify the set of relevant tools that can be used depending whether data or knowledge is desired.
1 Scientific Foundation

Data mining research can trace back to three distinct scientific disciplines: statistics, artificial intelligence, and machine learning.

While the term “data mining” did not gain widespread acceptance until the 1980s, and has been around in various forms. Its ancestry can be traced back through three distinct disciplines and are still very apparent in data mining research today: statistics, artificial intelligence, and machine learning (Chen and Chau 2004).

In statistical research, data mining evolved as a method to find the reasons behind relations. From statistics, we can find and measure the strength of a relationship (e.g., co-variance) between two variables from the data. But this statistical measurement by itself is unable to explain why the relationship exists or what possible impact it may have in the future (DataSoftSystems 2009). Data mining provides us with the tools to interrogate the data and gain further knowledge about the dependency relationships. It does so through an interactive, iterative and/or exploratory analysis of the data. From the statistical branch of the data mining genealogy, methods such as regression analysis, discriminant analysis and clustering became data mining tools (Data Mining Software 2009).

In regression analysis, data points are examined by fitting a line or polynomial to explain the most data while at the same time minimizing the fitting error. From a regression analysis, causal relationships can sometimes be discovered between dependent and independent variables. Most oftentimes, this type of analysis is used for prediction where if the observed trend holds true, estimates of the dependent variable’s value can be made by varying the values of independent variables.

Discriminant analysis is a classifier-based approach which seeks to best categorize the data based upon the combination of features that contribute to maximally separating the data. This analysis can not only identify the important classification features, but also provide insights into the relationship.

Clustering is another classifier method which seeks to partition the data into different sets based on their features. Clustering can take many forms including hierarchical clustering, or sub-partitioning of data, density-based clustering where item similarity can be measured, and clustering based on Euclidian distances where the distance between two clusters can provide a measure of similarity between them.

The second branch of data mining, artificial intelligence differs from statistics by applying a heuristic algorithm to the data. Heuristic-based approaches are experience-based techniques that can be computationally intense. It wasn’t until the 1980s that computers began to possess sufficient power to handle the complexities associated with heuristics. This approach attempts to balance out the statistics by applying a human problem-oriented perspective to problem solving based upon educated guesses or common sense. Imparting this problem-solving technique to a computer is a matter of learning from appropriate rules or cases. Heuristic solutions may not be perfect, however, the solutions generated are considered adequate.
Rule-based heuristics are a set of conditional statements that the computer tests in sequence in order to obtain a solution. This style of problem solving became popular during the 1980s in expert systems. Rule-based heuristics can derive new knowledge from the existing problem through inductive reasoning.

Case-based heuristics, or case-based reasoning, approaches problems based on solutions to past problems. This style of decision-making relies on having a sufficient breadth of historical cases and solutions to draw upon in order to create solutions by analogy.

The third branch, machine learning, uses algorithms to learn and induce knowledge from the data (DataSoftSystems 2009; Chen and Chau 2004). Examples of these algorithms include both supervised and unsupervised learning techniques such as genetic algorithms, neural networks, self-organizing maps and Bayesian methods. These techniques can iteratively identify previously unknown patterns within the data and add to our understanding of the composition of the dataset.

Genetic algorithms (GAs) are a stochastic optimal search technique based on evolutionary processes, where parallel solutions can evolve through crossover and mutation. In a GA, a number of initial solutions are created, tasked with the overall goal of finding an optimal solution. When one generation has finished, the solutions are then allowed to pair up and create offspring solutions for the subsequent generation. Not all solutions will participate in this procreation process as it is governed by chance. However, stronger solutions are given more chances than weaker ones to make it into the next generation. The solutions that have been selected then perform a cross-over to build offspring solutions and the parents are then removed. Also, as with evolutionary development, a small number of mutations are permitted which can sometimes have the effect of creating a much better solution by accident. This indeterminate process is then allowed to continue until a stopping condition is met which could be a preset number of generations, or more commonly the point where further generations do not appear to produce better results.

Neural networks are another machine learning technique that can learn to classify from data patterns. The most common form of Neural Network is the Back Propagation Neural Network (BPNN), which is a supervised learning technique that adjusts itself to the data patterns and can be used to make predictions from unseen data. BPNNs attempt to simulate human synaptic activity by modeling the brain’s mesh-like network of neurons and axons as perceptrons and weighted links. From adjusting the weighted links based on the training classes, BPNNs can recognize complex data patterns and learn to anticipate them.

Self-organizing maps (SOMs) are a special type of neural network that are used to visualize multidimensional data in lower dimensions, typically a two-dimensional plane. This visualization can help to identify trends or tendencies that can be clouded by higher orders of dimensionality.

Bayesian statistics considers the probability of states given the data. These probabilities include prior probability, a measure of the probability of a state condition being met before the introduction of data, conditional probability of seeing that data if the state is correct, the marginal probability of the data considering all states and
posterior probability that measures the belief in the state condition being met after seeing the data. This technique can then provide a measurable value of possible states given the data.

2 Traditional Data Mining Applications

Both structured and unstructured data can be “mined” for knowledge in different scientific, engineering, business, biomedical, public safety, and sports applications.

Data mining in a business environment is a common way of obtaining organizational intelligence. Traditionally, data mining within business has been concerned with capturing new knowledge within the data. This data could be somewhat structured yet raw, such as retail purchases, or be unstructured and difficult to computationally parse, as can be the case with textual data. Whether it be structured or unstructured, once the data has been parsed, cleaned and organized, data mining can take place.

Examples of data mining discovery on structured data routinely include using the customary grocery/retail illustrations. Beer and diaper sales have long shown a stronger than expected correlation in sales. While statistics by themselves can be used to identify these types of relationships, such as covariance or ANOVA-type analyses, the statistics cannot explain why the relationship exists, simply that it does. This is where data mining becomes an invaluable tool, by allowing us the answer the why type of questions and in the process gain new insights into the relationship. As it turns out, data mining the customer demographics of these sales discovered that men sent to the store (by the wives) to buy diapers, would also pick up a case of beer. Once discovered, some grocery stores changed their product layouts such that beer and diapers were located proximate to one another, in some cases within the same aisle, in order to increase impulse consumer purchases.

A second example of effective data mining usage in the retail sector comes from Walmart’s ability to track colds and flu more quickly and precisely than the US government’s Center for Disease Control (CDC). Walmart employs a sophisticated product inventory and reordering system that is primarily data driven. Using this product data, Walmart knows what the baseline sales of cold and flu remedies should be for their stores. When sales of these products exhibit a statistical spike in sales, a cold or flu outbreak has occurred within that store’s sphere of influence. Congregating the data together and adding in the geographic locations of their stores, Walmart can not only see the spread of colds and flu from store to store, but also can predict future affected stores and therefore stock their shelves appropriately.

As a third example of structured data mining usage, and an example outside of the retail arena, police reports are a good example of well structured documents that can
be used to detect criminal aliases. When criminals do not want to be positively identified, they will provide false information such as name, address, date of birth, etc. In order to combat the introduction of false information, police will routinely ask for the information multiple times. A careless criminal may not remember their previous lie and get caught by this tactic. Criminals that know this tactic, will provide information that is close to the real data so that they can remember their falsehood and perpetrate it successfully. This misdirection can lead to multiple criminal aliases which can be laborious for an officer to detect. By introducing textual tools that comb through police reports and look for data that is close, but slightly off, the officer’s workload is reduced and criminal deception can be discovered (Chen 2006).

Unstructured data, and notably textual data, is more difficult to manage. Processes must be put in place to properly extract the data from these sources. To do this, there are two major schools of thought; a template-based approach where data is identified and pigeon-holed into particular fields using a template, or a lexical/syntactic or semantic analysis of the textual structure or meaning of the text. Template-based approaches work great when there is some form of order to the text. Sales receipts, product invoices and shipping documents all fit this definition because their defined structure allows a textual parser to effectively extract the data. Memos, letters and newspapers articles do not conform to template-based methods and require a different type of analysis. In a lexical/syntactic analysis, the structure of the text is examined. Terms are categorized by their parts of speech using either a lexicon or dictionary, or identified by the syntax of the sentence. Examples of using this type of analysis include predicting stock prices from financial news articles, identifying specific cancer gene pathway associations in medical abstracts and identifying unknown authors by their writing style tendencies (Chen 2006).

In predicting stock prices from financial news articles, terms within news articles give us clues as to the probable price movement of a company. Terms such as “factory exploded” or “workers strike” typically indicate a downward price movement while “unexpected gains” and “earnings surprise” may indicate upward trends. By analyzing the article terms against price movements, stock prices can be predicted with better than chance accuracy immediately following the release of the news article. Another example includes identifying relations in cancer gene pathways from medical abstracts. Oftentimes, medical researchers will focus their attention on particular subsets of cancer research. Whether it be an enzyme reaction, or observed effects of different genes, researchers write about their experiment and move on. The problem is, with so many researchers focusing on the minutia of cancer, there isn’t a lot of work on the bigger picture, or putting the dots together and getting a more complete picture of the disease. However, it is quite difficult to achieve this, because it would require a medical researcher to become familiar with a majority of current research as well as possess an ability to put it all together. This problem is well suited to textual data mining and machine learning techniques that can identify the gene pathways described in medical abstracts and find previously undiscovered links. A third example includes discovering the identity of authors based upon previous writing samples. All individuals have certain tendencies within
their writing styles. Whether it be how a sentence is constructed, particular word choices or punctuation usage, these indicators can be used to identify the similarities between an unknown author and a body of known ones (Chen 2001, 2006).

A semantic analysis instead takes aim at the meaning of the terms by applying either a hierarchical framework, as is the case with WordNet and Cyc, or looks at the context of sentence for clues to problems such as word-sense disambiguation. Applications that employ Cyc, a type of hierarchical encyclopedia for computers, can gain additional insight not otherwise obtained. As an example of using this technology, an application that came across the phrase “lemon tree” could query Cyc and traverse the hierarchy upwards to discover that lemon tree is a type of citrus tree which is a type of deciduous tree which is a type of tree which is a type of plant. The application could also seek a more detailed explanation by traversing the hierarchy downwards and discovering that Lisbon lemon is a specific instance of a lemon tree. In word-sense disambiguation, some terms may be ambiguous to the computer. Bank is one such example, where the term could mean a financial institution or an adjoining land area proximate to a waterway. By analyzing the terms surrounding the ambiguous term, clues can be discovered to distinguish its meaning.

3 Deriving Knowledge

At each level of the Data-Information-Knowledge-Wisdom hierarchy, meaning can be obtained through filtering and analysis, and data can tell a story and explain the observed phenomenon.

Returning to the DIKW hierarchy, each level can provide specific meaning to describe the dataset. This meaning is obtained through filtering of relevant material and abstracting it in such a way that the data begins to tell a story and explain the observed phenomenon.

Data are the observable differences in physical states (Boisot and Canals 2004) that are acquired from stimuli and examination of the world around us. By themselves, data are generally overwhelming and not entirely useable. In our framework, data can be thought of as all of the individual events that took place in the sports match. If applied to baseball, this data could contain a record of pitch sequences, at-bats and defensive moves which by themselves provide little interest or value.

To be of practical value, data needs to be transformed by identifying relationships (Barlas et al. 2005) or limited to only that which is relevant to the problem at hand (Carlisle 2006). This data transformation produces information which can be characterized as meaningful, useful data (Bierly et al. 2000). Returning to our baseball example, information could focus on the pitch sequences by a particular pitcher or batting sequence of a certain batter. Although information is not very
useful at this stage, abstracting it to the next level of the hierarchy, knowledge, can provide additional meaning by identifying patterns or rules within the data.

Knowledge is the aggregation of related information (Barlas et al. 2005), that forms a set of expectations or rules (Boisot and Canals 2004) and provides a clearer understanding of the aggregated information (Bierly et al. 2000). At this level of the DIKW hierarchy, rule-based systems can be developed to aid individuals in both expanding their own knowledge while also providing a benefit to the organization (Alavi and Leidner 2001). Using our baseball example again, analysts can evaluate the pitching information of a particular player and look for tendencies or expectations in the types of pitches thrown. Data mining can then be defined as the pursuit of knowledge within the data.

Precise definitions of data, information and knowledge are still a matter of debate within the Knowledge Management community. So is it with the final level of the DIKW hierarchy, wisdom. Wisdom can be viewed as a grasp of the overall situation (Barlas et al. 2005), that uses knowledge and knowledge alone (Carlisle 2006) to achieve goals (Bierly et al. 2000). In our baseball example, we have several disparate pieces of knowledge as well as an ultimate goal. From data mining we can obtain knowledge of the types of pitches to be expected, knowledge of effective strategies to approach specific types of pitches and an overriding goal that a successful at-bat can help win a game. Putting all of this knowledge together into wisdom; the batter has a chance to positively influence the game in their favor if the observed pitching data pattern holds and the batter is able to use it to his advantage. Wisdom resides in the capabilities of cognition and human understanding (Carlisle 2006), as a computational approach is currently difficult to capture (Barlas et al. 2005).

4 Questions for Discussion

1. Can you see any circumstance where data mining of sports data may not be beneficial, and if so, is it because of problems with performance metrics or data mining itself?
2. Choose a sport and analyze how the three branches of data mining (statistics, artificial intelligence, and machine learning) could be used to derive knowledge.
3. Is there any particular sport which lends itself better to one particular branch of data mining, and why?
Sports Data Mining
Schumaker, R.P.; Solieman, O.K.; Chen, H.
2010, XIV, 138 p., Hardcover