Mining Ice Hockey: Continuous Data Flow Analysis

Adam Hipp
Applied Computational Intelligence Laboratory
University of Cincinnati
Cincinnati, Ohio 45220
hipp.adam@uc.edu

Lawrence J. Mazlack
Applied Computational Intelligence Laboratory
University of Cincinnati
Cincinnati, Ohio 45220
mazlack@uc.edu

***Abstract*—Ice hockey is relatively under computationally analyzed. Possibly this is because ice hockey is a continuous flow game with relatively few major events (goal scoring) while most of the other games that have been data mined can be described as being a series of clearly bounded events. This work describes needs of data mining ice hockey statistics to quantify the contribution of individual hockey players to team success. Large databases of ice hockey statistics for the collegiate and professional levels can be accessed to perform this work. The goal is to use ice hockey statistics and computational methods to help make personnel decisions at both the coaching and franchise management levels. This type of work has the potential to encourage new avenues of sports statistics research, as well as statistical research and data mining, in general.**

Keywords-data mining; continuous flow analysis; ice hockey

I. Objectives and Overview

Data mining techniques have been developed to take large bodies of statistics and reduce them to interesting patterns and relationships using computers much faster than a human could on his own. Data mining techniques have been successfully applied to many types of large databases, and an emerging application of data mining is sports statistics. Sports generally lend themselves well to data mining since most sports contain large amounts of statistics that have been kept over a long period of time.

The purpose of this work is to data mine ice hockey statistics to quantify the contribution of individual hockey players to team success (i.e., winning). Large databases of ice hockey statistics for the collegiate and professional levels can be accessed to perform this work. The long term goal is to use ice hockey statistics and computational methods to help make personnel decisions at both the coaching and franchise management levels. For example, applications of this work could be used to help a coach decide which players on his team should receive more playing time, or to help a team executive decide who on his team should be traded and for whom from another team.

The central hypothesis is that currently available ice hockey statistical databases, which contain statistics compiled for individual players, can be computationally used to aid in making effective decisions about ice hockey players. The rationale behind this hypothesis is that computational approaches to player statistics in other sports in a similar fashion have been successful in the past, and as stated, there is a large amount of hockey statistics available for analysis. There are two specific aims:

- **Create a computer model that takes ice hockey statistics as an input and scores each player's contribution to their team**. The working hypothesis is that currently available ice hockey statistics can be used to effectively compare players to one another on a single, useful, and unbiased scale.
- **Create a more robust computer model that takes multiple players' statistics into account to quantify how effective multiple players are together.** The working hypothesis is that currently available ice hockey statistics can be used to score a "squad" contribution level on a single, useful, and unbiased scale.

While data mining methods have been applied to some sports with success, very little research has been done specifically on ice hockey. The results add to the emerging field of sports data mining.

II. Significance

Data mining and statistical analysis has been applied to other major sports with some success, but very little research has been done with ice hockey statistics. The argument could be made that this is because ice hockey does not lend itself to this type of analysis as easily as sports such as basketball, baseball, football, and soccer. The difference between ice hockey and these other games can be identified by the internal game flow. In baseball and football, action is broken into a sequence of separate plays; the result of what each player does in a particular play is easily quantified, for example: bases accumulated on a hit in baseball or yards gained on a single carry by a running back in football. In basketball, plays are not as broken up, but scoring is frequent, causing more statistics to be accumulated for each player throughout the course of the game. Hockey is a low scoring, relatively constant flow game.

There are a significant amount of hockey statistical categories, even if the number of categories is less than the other games. The primary statistics a player accumulates are goals, assists, points (goals + assists), and +/- (difference between team goals scored and team goals allowed while the player is on the ice). Other lesser known statistics have been compiled as well. The website of the National Hockey League (NHL), NHL.com, keeps detailed player statistics dating back to 1997, and appends these basic statistical categories with penalty minutes, power play goals, shorthanded goals, game winning goals, game tying goals, overtime goals, shots, shooting percentage, time on ice, shifts per game, and face-off win percentage. There is

 ISBN: 978-1-61208-162-5

potential in the accumulated available data to quantify player contributions using more complexity than just goals and assists. In depth statistical analysis of a given set of sports statistics can lead to the creation of new statistical categories, causing useful analysis to build upon itself.

Based on the successful application of statistical techniques and data mining to other sports and the untapped potential of ice hockey statistics, the significance of the gap in knowledge of applying these techniques to ice hockey is made clearer.

This work is expected to improve the decision making abilities of coaches, team executives, and other people involved in management of an ice hockey team. There are many types of benefits that can come out of the new knowledge generated by this work.

One group of people who will benefit from the work is ice hockey coaches. The largest beneficiaries would be hockey coaches at the highest level, professional or international team coaches. This is because the most statistics and interest are available at the professional level. The results of this work can be used to help a coach decide which players should receive more playing time, or when different players can be ideally used in different situations. The results will also help a coach identify players who are undervalued on his roster or what players would benefit most from what type of development. Over time, a coach may learn to identify traits in a player that may have been suggested by the statistical analysis, which can improve a coach's talent and instincts.

A second group of people who will benefit from this work are team executives If this analysis identifies a player on another team who is undervalued by his team, a team executive may use this information to propose a trade for the undervalued player without having to give up too much for a potentially good player. This work can also help team executives decide what players to draft from colleges and other lower leagues without having to rely purely on scouts who can only investigate a limited number of players. This work can help teams discover players who might otherwise not have been given a chance.

Another benefit of this work is the benefit to the followers of ice hockey. By having better players and better teams taking the ice every day, the quality of the game will be improved, which will make the game more enjoyable for the fans. Beyond this, many fans of sports enjoy hearing statistical analyses of the games they follow, and this work will open the doors to new types of hockey analysis. More interesting statistics will help make ice hockey even more interesting and generate more avid followers of the game.

III. BACKGROUND

There has been some research into applications of statistics and data mining to sports. This research has approached the problem of modeling sports situations and performance in a variety of ways. One vein of sports research that can be found in the literature is research which attempts to statistically model in-game situations. This can be most commonly seen with soccer, since, like hockey, soccer is a constantly moving, low scoring game, leading to less statistics to be generated.

For example, Wang and Wang [1] used the Apriori algorithm [2] to determine association rules between different types of technical movements in soccer. The idea behind this research was to determine what the most common chains of ball movements that occur during a particular game are. This type of research can be applied to the ball movement patterns of a particular team to determine what kind of defense to play against the team, or to help players predict what the opposing team's next move may be after something occurs based on common occurrence chains determined from application of this research. The only thing required would be for someone to watch the opposing team play and keep a record of when each technical ball movement occurred throughout the game, generating a long chain that represented the entire game for the team. By watching many games, a database of ball movement can be created and then mined using the research as presented.

A similar approach was taken by Chai [3] to mine patterns from data of soccer using serial data, pass time, and ball possession time. Serial data refers to the pattern of which player possesses the ball in a given chain of possession for a team. Chai applied data mining techniques to the serial data to determine what chains of possession between players happened often, incorporating both teams with specific players into each pattern. This work can be used to show what trends of ball possession happen when a specific team is in control. Also, it can help to identify weaknesses within that team's strategy by incorporating the players on the opposing team. If a particular chain of possession for a team was often broken up by a particular player on the opposing team, a future opponent can use that information to adjust its strategy.

Beyond just the actual patterns of player possession determined, the pass time and ball possession time was used to create a time series model of a particular game, which led to other association rules about each interesting chain of player possession which gave some insight into why events occurred. For example, a chain of four specific players may end in a steal by the opposing team often, but the reason why would be up to speculation from that information alone. However, by knowing that the specific chain of players led to longer pass times, it can be inferred that these players tend to make long, dangerous passes to one another, which can be used to help strategize against that team in the future. [3]

Nunes and Sousa [4] applied different data mining and data visualization techniques to available data relating to the European championship soccer matches over the history of the tournaments. Using in game data such as cards given, goals scored, and substitutions, these techniques were able to confirm the types of trends and patterns one may expect for a soccer match. For example, the research showed that substitutions begin occurring towards the end of the first half, with very few substitutions early in the match. Since when a player is substituted out in a soccer match, he is done for the rest of the game, it makes sense that substituting early in the match would be uncommon.

 ISBN: 978-1-61208-162-5

Halftime is often a good time to make a substitution since it gives coaches and players time to refocus their strategy with the break.

Nunes and Sousa [4] applied the Apriori algorithm to match specific information from the tournament's history in the hopes of finding interesting and unexpected association rules, but the results of this portion of the research did not yield interesting results. Instead, trivial association rules were found. The research went on to apply classification data mining techniques to matches in different countries but it was again unable to find many useful classifications in terms of predicting the outcome of soccer matches. Data visualization proved to be the most useful of the statistical techniques applied to the data in this research.

There have been various levels of success applying game-model statistical research to soccer matches. Soccer and ice hockey share many traits in terms of how the game can be modeled, and one avenue of ice hockey statistical research that may be worth pursuing is the one taken for soccer research shown here. However, that is not the type of data mining for ice hockey that is described here. Data mining that makes use of a greater number player and team statistics in a more generalized fashion, which is more similar to the work described here, can be found in the literature.

Maheswari and Rajaram [5] decided to apply data mining techniques to cricket data, noting that cricket lends itself to large accumulations of data which cannot be completely analyzed by a human in a reasonable amount of time. The approach in this work straddles the line between game flow modeling, similar to the soccer research just presented, and purely accumulated statistical modeling. In cricket, one team bowls the ball towards the batter, who is a member from the other team. The batter is on offense and the bowler is on defense. Play can be broken up by each bowl, where the bowler bowls the ball and the hitter hits it. A sequence of events will occur which may or may not result in runs being scored, and then the process repeats with a new bowl. Since cricket is a game where many short plays occur, similar to baseball, association between statistical events is much clearer to infer. This work uses data mining algorithms to find association rules between different statistical categories in cricket for different players. This type of information can be used to find a player's strengths and weaknesses to different approaches from the opposition. For example, the type of bowl to which a hitter is weak could be discovered through an application of this work helping the defensive team strategize as to how to bowl to that particular player.

Recently, Wang, Jie, and Zeng published a paper related to their work on development of an interactive baseball and softball statistical gathering and analysis computer program [6]. Fast and Jensen [7] applied data mining techniques to broad NFL team statistics, such as wins, losses, points scored and points allowed, as well as the history of NFL coaching performances and associations between coaches. Coaching associations comes in the form of who was an assistant under whom, to give an idea of who learned from successful coaches and went on to become successful himself. Based on statistical team success and coaching associations, the research was able to show that the history of the coaching staff for a given team had a strong association to whether that team made the NFL playoffs in the given year. This type of research could be used to help identify up and coming or undervalued coaches quickly, so a franchise or school would move in to acquire this coach before other teams or schools start to take notice. Conversely, this research could also be used to show what coaches may be given more credit than they deserve.

Romer [8] applied statistical analysis and dynamic programming to model NFL football game situations with the goal of determining optimal strategy on fourth-down plays. The generally accepted strategy in football is for the team with the ball to punt in many given situations when the team has the ball on fourth down. Romer's analysis shows that conventional knowledge may be approaching the game too conservatively, meaning that teams should be more willing to "go for it" on fourth down to try to keep possession of the ball if they are able to gain enough yards.

There has even been sports research from a more media-driven point of view, such as the research by Dao and Babaguchi [9]. This research uses a temporal representation of small events during the course of a sporting event to predict what type of event will occur next. The results of the research could be useful for media coverage of sporting events, to help a system automatically identify when a replay was likely to be displayed to viewers, or for the system to understand what kind of event is happening on the field based on the positioning of cameras at the current time and moments leading up to the current time. This type of automatic detection could be used to improve the speed and quality of media coverage of sporting events. Since this type of detection would be nearly impossible to have perfect results, the best application would likely to let the system automatically suggest to an interactive user what the next step in coverage to take should be, which could speed up the decision making process within the media and thus enhance the viewer experience.

An application of data mining takes a similar approach to ours was done by Smith, Lipscomb, and Simkins [10]. In their work, the goal was to use data mining of individual baseball pitcher statistics to predict the winner of the annual Cy Young award. The Cy Young award is given to the pitched voted to be the best pitcher in each of the two leagues that make up Major League Baseball. The application in this research chose the top 10 starting pitchers in terms of wins and top 10 relief pitchers in terms of saves as candidates for the Cy Young in the given season. The individual statistics available for each candidate are then run through a data mining algorithm that is a Bayesian classifier. Applying the classifier to data from 1967 through 2006, the correct pick of Cy Young winner from each league happened frequently.

In the 1990's, IBM developed a program called Advanced Scout [11] that was created to "seek out and discover interesting patterns in game data." Advanced Scout was a specific data mining application for NBA basketball, and was distributed to many NBA teams during the 1990's.

 ISBN: 978-1-61208-162-5

Using input data from a given game of basketball between two teams, which includes very specific input data such as "who took a shot, the type of shot, the outcome, any rebounds, etc." The data is also temporal, showing when each data point occurred during the game. The IBM software is then able to apply data mining algorithms to this data and find interesting and useful patterns and associations. Since the data is temporal, video can accompany a game analysis, so when a particular pattern is shown to occur at a time point, a coach can then pull up the video to actually see what the statistics are saying. This type of interaction can be very valuable to a coach, helping him to learn to spot important patterns and behaviors on his own, as well as to better understand those identified by the data.

Some of the general data mining queries involve "either field goal shooting percentage to detect patterns related to shooting performance, or possession analysis to determine optimal lineup combinations." Advanced Scout takes an approach similar to the cricket application by Maheswari and Rajaram [5], but with more detailed input data and more rigorous analysis.

Another broader-focus type of sports data analysis is known as "sabermetrics." The beginnings of sabermetrics are well explained by Schumaker [12]: "In 1977, Bill James began publishing his annual *Bill James Baseball Abstracts*. These abstracts were used as his personal forum to question many of the traditional baseball performance metrics… James continued to publish his annual compendium of insights, unorthodox ranking formulae and new statistical performance measures which he called *sabermetrics*." Baseball is a statistics rich game, and it is now accepted that rigorous statistical analysis is necessary to remain competitive at the professional franchise level of baseball. However, James' sabermetrics were not even strongly considered by decision makers in Major League Baseball until 2002, when the Oakland Athletics began to incorporate some of James' ideas. The A's used some of James' performance measurements to help decide whom to draft, and the strategy marked a turnaround for the A's franchise. The A's general manager, Billy Beane, "discovered that by carefully selecting players in the draft, the A's could lock-in players that were oftentimes overlooked by other clubs into long contracts that paid little money and thus develop this into a strategy to compete with larger payroll teams. It was simply a matter of picking the right players, which sabermetrics could make easier."

The traditional baseball statistics are categories like runs batted in (RBI), hits, singles, doubles, triples, home runs, and stolen bases for field players, and categories like earned run average for pitchers, which takes the number of earned runs a pitcher gives up and normalizes it by 9 innings to give a measure of the average number of runs that pitch would give up in a normal nine inning game. James came up with different categories that helped to better quantify player performance, such as On-Base Percentage, which determines what percentage of plate appearances for a batter results in the batter reaching base, and strikeout to walk ratio for pitchers.

Schumaker [12] explained how Dean Oliver took a similar approach to basketball statistics as James did to baseball beginning in the 1980's. Oliver took more of a team-based statistical approach, but still was able to make contributions to analysis of basketball by helping to identify player contribution and even measure how well players worked with one another on the court.

As the success of analysis such as that of James and Oliver began to be noticed, a sudden revolution in sports data analysis began and is continuing today. A sabermetrics type of analysis is beginning to be applied to football as well, which has the benefit of many types of statistics easily able to be accrued due to the stop-and-go nature of the game, but the drawback in that there are only 16 games in an NFL season, while there are 162 games in a baseball season and 82 games in an NBA season. As a comparison, there are 82 games in an NHL ice hockey season as well. [12]

There have been plenty of statistical computer tools that have been created to access and analyze the large amount of sports statistics available. Schumaker [12] provides a short summary of different programs, some of which are for a specific sport, while others can be used for many different games. One example is Digital Scout, which is an adaptable piece of software that can be used for record keeping and creating custom reports, "such as baseball hit charts, basketball shots, and football formation strengths." Another is SportsVis, which creates graphical representations of sports data quickly to help people uncover trends or problems.

Schumaker [12] delved into research tools and methods created for a large variety of sports, but ice hockey information is noticeably sparse. Comprehensive statistical data about players and teams in hockey can be found in two major online resources are supplied by the NHL [13] as well as an outside, independent resource [14]. Interactive graphics that can be used to identify where events occur on the ice and when can be found are also supplied by the NFL [13]. Similar to basketball, a shot-taking map can be drawn on a graphical representation of the ice/court, and the visual data can infer trends or patterns about players or a team [15].

IV. DATA

There are two specific objectives of the work. The first objective is to create a computer model that takes ice hockey statistics as an input and scores each player's contribution to his team. The second objective is to create a robust computer model which takes multiple players' statistics into account to quantify how effective multiple players are together.

To accomplish both of these objectives, a large amount of data needs to be accessed. NHL.com has comprehensive statistics reaching back to the 1997-1998 season for every player, including Games Played, Goals, Assists, Points, +/-, Penalty Minutes, Power Play Goals, Short Handed Goals, Game Winning Goals, Overtime Goals, Shots, Shooting Percentage, Time on Ice per Game, Shifts per Game, and

 ISBN: 978-1-61208-162-5

Face-off winning percentage. The website also has game-by-game statistics and team statistics.

Since the first goal is to score a player's contribution to team success, both the individual player statistics and team statistics will be important to consider. For example, if one player has scored many more goals then another, but the high-scoring player is on a team with many more losses than the lower-scoring player, that doesn't necessarily mean that the higher-scoring player is a better contributor to team winning. There could be something else the player does that makes it more difficult for his team to win, which may even be related to his higher scoring. If the higher-scoring player has a tendency to play too aggressively, he will score more goals, but he'll also put his teammates in more shorthanded situations, either through penalties or by trying to take the puck too far on his own, causing him to turn the puck over more often and lead to breakaways for the opposing team. The basic idea is to use player statistics and the team statistics accumulated at the same time. With a large database to work with, the hope is to mine interesting association rules. However, the approach for mining these rules will not be straightforward. In fact, players may need to be grouped differently depending on their position. For example, penalty minutes accrued by a center may have an association with accumulating losses for a team, but penalty minutes accrued by a defender may have an association with accumulating wins for a team. This is a possibility due to the different nature of the positions. A center's job is to win face-offs and score goals; a defender's job is to stop the opposing team from scoring. Generally, a rougher, more physical player is a desirable attribute in a defender, and those types of players will tend to accumulate more penalty minutes. Again, this is speculation, and quite the opposite may result from the analysis.

Once a variety of data segmentation and approaches to mining association rules between player statistics and team wins have been run, the intention would be for a comprehensive formula that scores each player's contribution to team success based on their statistics. The formula will be regardless of who the player is or what team the player is on. This way, the value of two players from very different backgrounds, including position differences and team history, can be compared directly in an unbiased way. If an effective formula can be derived, it could be used to identify a player who makes strong contributions to his team but has been stuck on unsuccessful teams due to the players around him, which could be used to the advantage of both the team who has the player and a different team who would want to acquire him.

To accomplish the second objective, a similar approach is taken as with the first, using pairs of defenders and lines of forwards, where a line consists of two wings and a center. The groups of players to be used for creating the association rules will be chosen by researching commonly used lines on teams throughout the time period covered by available data. This is because randomly choosing players to create a "line" for each team would work statistically, but could cause misleading results since the players may not necessarily work together. The goal is to derive association rules that may have a form like "If defender 1 and defender 2 both have an average of more than 2 penalty minutes per game with an average time on ice of at least 20 minutes per game, then wins occur."

Similar to the first process, if a useful set of interesting association rules can be derived, the next step would be to develop a mathematical formula using the statistics of a pair or set of three players and "score" the group of players. If actual scoring cannot be accomplished, at least a set of player types that create an idea pair of defenders or line of forward could be suggested with statistical support. With the complex level of analysis required for combining players, even suggesting what "groups of statistics" for each of two or three players on a line could be very valuable in constructing lineups for a game or figuring out who to draft or acquire for a team, based on the current roster.

VI. CONCLUSION

There is sufficient precedent to show the success and benefits of sports statistics and data mining analysis. Ice hockey is relatively under computationally analyzed. Possibly this is because ice hockey is a continuous flow game with relatively few major events (goal scoring) while most of the other games that have been data mined can be described as being a series of clearly bounded events. This work described the needs of data mining ice hockey statistics to quantify the contribution of individual hockey players to team success. Large databases of ice hockey statistics for the collegiate and professional levels can be accessed to perform this work. The goal is to use ice hockey statistics and computational methods to help make personnel decisions at both the coaching and franchise management levels. Many people from the walks of academia and business, as well as hockey enthusiasts, will benefit from the results of this research. This type of work also has the potential to encourage new avenues of sports statistics research, as well as statistical research and data mining, in general. It is an important societal goal to find an effective way to draw more people into studying fields involving math and science.

REFERENCES

[1] B. Wang and L. Wang, "Research of Association Rules in Analyzing Technique of Football Match," presented at Second International Conference on Power Electronics and Intelligent Transportation Systems, 178-180, 2009.

[2] R. Agrawal and R. Srikant, "Fast Algorithms For Mining Association Rules," presented at 20th VLDB Conference, 487-499, 1995.

[3] B. Chai, "Time Series Data Mining Implemented on Football Match," *Applied Mechanics and Materials*, vol. 26-28, pp. 98-103, 2010.

[4] S. Nunes and M. Sousa, "Applying Data Mining Techniques to Football Data from European Championships," presented at Conferência de Metodologias de Investigação Científica (CoMIC'06), 4-16, 2006.

[5] P. Maheswari and M. Rajaram, "A Novel Approach for Mining Association Rules on Sports Data using Component

 ISBN: 978-1-61208-162-5

Analysis: For Cricket match perspective," presented at 2009 IEEE International Advance Computing Conference, 2009.

[6] G. Wang, S. Jie, and F. Zeng, "Design and Realization of Baseball and Softball Match Data Analysis Information System," *Advanced Materials Research*, vol. 187, pp. 353-357, 2011.

[7] A. Fast and D. Jensen, "The NFL Coaching Network: Analysis of the Social Network Among Professional Football Coaches," presented at AAAI Fall Symposia on Capturing and Using Patterns for Evidence Detection, 2006.

[8] D. Romer, "It's Fourth Down and What Does the Bellman Equation Say? A Dynamic-Programming Analysis of Football Strategy," 9024, 2003.

[9] M. Dao and N. Babaguchi, "Sports Event Detection using Temporal Patterns and Web-casting Text.," presented at First ACM Workshop On Analysis And Retrieval Of Events/Actions And Work Flows In Video Streams (AREA '08), 2008.

[10] L. Smith, B. Lipscomb, and A. Simkins, "Data Mining in Sports: Predicting Cy Young Winners," *Journal of Computing Sciences in Colleges*, vol. 22, pp. 115-121, 2007.

[11] I. Bhandari, E. Colet, J. Parker, Z. Pines, R. Pretap, and K. Ramanujam, "Advanced Scout: Data Mining And Knowledge Discovery In NBA Data," *Data Mining and Knowledge Discovery*, vol. 1, pp. 121-125, 1997.

[12] R. Schumaker, O. Solieman, and H. Chen, "Research in Sports Statistics," *Sports Data Mining*, pp. 33-50, 2011.

[13] http://www.nhl.com/

[14] http://www.hockey-reference.com/

[15] H. Chen, R. Schumaker, and O. Solieman, "Data Sources for Sports," *Sports Data Mining*, pp. 25-31, 2011.

 ISBN: 978-1-61208-162-5