Data Mining and Knowledge Discovery, 1, 121–125 (1997)

Brief Application Description
Advanced Scout: Data Mining and Knowledge Discovery in NBA Data

INDERPAL BHANDARI isb@watson.ibm.com
EDWARD COLET ecolet@watson.ibm.com
JENNIFER PARKER jparker@watson.ibm.com
ZACHARY PINES s5pines@watson.ibm.com
RAJIV PRATAP c1rajiv@watson.ibm.com
KRISHNAKUMAR RAMANUJAM c1kk@watson.ibm.com
IBM T.J. Watson Research Center

Editor: Gregory Piatetsky-Shapiro

Received April 30, 1996; Revised August 15, 1996; Accepted August 15, 1996

Abstract. Advanced Scout is a PC-based data mining application used by National Basketball Association (NBA) coaching staffs to discover interesting patterns in basketball game data. We describe Advanced Scout software from the perspective of data mining and knowledge discovery. This paper highlights the pre-processing of raw data that the program performs, describes the data mining aspects of the software and how the interpretation of patterns supports the process of knowledge discovery. The underlying technique of attribute focusing as the basis of the algorithm is also described. The process of pattern interpretation is facilitated by allowing the user to relate patterns to video tape.

Keywords: data mining, knowledge discovery, attribute focusing, basketball, NBA

1. Introduction

Advanced Scout (AS) software seeks out and discovers interesting patterns in game data. With this information, a coach can assess the effectiveness of certain coaching decisions and formulate game strategies for subsequent games. In the 1995-96 season, AS software had been distributed to sixteen of the twenty-nine NBA teams. While some of the teams were at the early stages of evaluating it's use, others (i.e. New York Knicks, Orlando Magic, Seattle Supersonics) quickly integrated the software into their game preparation and analytical processes. The positive feedback received from coaching staffs indicated that it's been a valuable tool. Bob Salmi (while at the NY Knicks), likened it to having another coach on the team. It has also been well received by the NBA because it contributes to improving the quality of play, which provides additional value to fans of the game (McMurray, 1995; Sterba, 1996).

2. Data collection

The raw data from NBA games is initially collected using a specialized system designed for logging basketball data. Data include who took a shot, the type of shot, the outcome, any

rebounds, etc. Each action is associated with a time code. At the end of each game, the data are uploaded and stored on an electronic bulletin board. Any team can access and retrieve the data of any other team from this billboard. A copy of the data must be downloaded into AS for analysis.

3. Data pre-processing: Cleaning, transformations, and enrichment

After the data have been downloaded, AS performs a series of consistency checks to ensure that the data are as accurate as possible before any analysis occurs. There are various aspects to data cleaning (Redman, 1992; McDonald and Celko, 1995). In AS, consistency checks are designed to detect errors made during the data collection process. A data error is a missing action or an impossible event. Corrections are made using a rule base, and/or with the input of a domain expert (typically a coach). For example, if two shots appear to be taken in quick succession without anyone credited with a rebound, the program will assume that the person taking the second shot also rebounded the ball. If needed, corrections can be verified via video tape.

After the consistency checks,the data are transformed and reformatted. This is to facilitate a coach's inspection of raw data and to define an appropriate unit of analysis that is consistent with their perspective. AS reformats the raw data of discrete events into a play-sheet - a standard form in which an event description and time are listed sequentially. Since coaches are very familiar with the play-sheets, they can quickly examine the discrete events of a game. From coaches' input, we discovered that the appropriate unit of analysis is an entire possession, composed of the actions and sub-events that precede a shot attempt, rather than the elementary discrete events. Therefore additional transformations are made to group events into possessions.

Data enrichment refers to the use of additional information to add value to analysis. Data is enriched through by inference rules and additional data entry. The role of each player on the court (e.g. power forward, 1-guard, etc.) is inferred by AS based on information in a player-role table. These inferences allow useful analyses of player-role relationships. Often, the plays a team decides to use are related to the players and their roles. Every team has a set of plays and many of a team's possessions are based on a specific play. Analyses of the circumstances surrounding the success or failure of their plays is therefore important. Because a team's plays are confidential and unknown to those logging the data, the play call information is not part of the raw data. The play call information is entered by the coach of the team, drawing from a separate, and confidential data source.

4. Data mining

Data mining can be viewed as the automated application of algorithms to detect patterns in data. (Fayyad, Piatetsky-Shapiro, Smyth, 1996). In AS, a coach can initiate a general data mining query in which the program will automatically search for interesting patterns for either the home or away team using either field goal shooting percentage to detect patterns related to shooting performance, or possession analysis to determine optimal lineup

combinations. Subsequent analyses may include more specific queries in which other attributes and conditions are specified.

The algorithms that underlie the data mining aspects of AS use a technique called Attribute Focusing (AF), (Bhandari, 1995). An overall distribution of an attribute is compared with the distribution of this attribute for various subsets of the data. If a certain subset of data has a characteristically different distribution for the focus attribute, then that combination of attributes, (the conditions that define the subset) are marked as interesting. Early applications of this technique were applied to software process engineering (Bhandari, 1993) and it has been extended as the data mining technique used in AS.

In the AF algorithm, an "interesting event" can be described more formally as the following:

An "event", E is a string $E_n = x_1, x_2, x_3, \ldots, x_n$; in which x_j is a possible value for some attribute and x_k is a value for a different attribute of the underlying data. E is interesting to the extent that x_j's occurrence depends on the other x_i's occurrence. The interestingness measure used by AS is the size $I_j(E)$ of the difference between: (a) the probability of E among all such events in the data set and, (b) the probability that $x_1, x_2, x_3, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n$ and x_j occurred independently. A first condition of Interestingness exists only if $I_j(E) > D$; where $D =$ some fixed threshold. When set at an appropriate level, this removes "false positives". A second condition of Interestingness depends on finding an optimal number of attribute values, n, formally described as $I_j(E_n) > I_j(E_{n-1})$; and $I_j(E_n) >= I_j(E_{n+1})$; where $E_n = x_1, x_2, x_3, \ldots, x_n$.

So, AF seeks to eliminate all but the most interesting events by keeping E only if the number of attribute values, n, is optimal: eliminate one or more x_i's, and I_j decreases, include one or more new x_i's to the string and I_j gets no larger. The convergence to an n removes patterns (an E_{n-i}, or E_{n+i}) which are less interesting than E_n, and have information already contained by E_n. Consequently, the user need not have to drill-down or drill-up from a highlighted pattern because the event descriptions returned by the algorithm are at their most interesting level.

5. Interpretation and knowledge discovery

The results of data mining are presented to the user in two forms - a text description and a graph (omitted here). Automatically generated text describes the patterns. An example reports,

When Price was Point-Guard, J.Williams missed 0% (0) of his jump field-goal-attempts and made 100% (4) of his jump field-goal-attempts. The total number of such field-goal-attempts was 4. This is a different pattern than the norm which shows that: Cavaliers players missed 50.70% of their total field-goal-attempts. Cavaliers players scored 49.30% of their total field-goal-attempts.

The objective of such a presentation is to ensure that the results are easily understood by a coach. (A point-guard is a basketball term that refers to the player responsible for bringing the ball up the court and directing the offense). The text presentation also offers a suggestion as to why the particular pattern is interesting - explicitly pointing out the ways that this particular pattern deviates from an expected norm - in essence presenting an initial

argument and easily interpretable justification of Interestingness. A subsequent requirement of knowledge discovery would be for the domain expert to determine the underlying cause of this pattern. (What was revealed in this particular case was that when the Knicks put two players on Price, he successfully found Williams unguarded for a jump shot).

The process of interpreting patterns represents knowledge discovery, and traditionally requires activity on the part of a domain expert. In AS, pattern interpretation is facilitated by providing the user with several opportunities for further interactive analysis to gain additional contextual information. For example, if play calls designed for Patrick Ewing are less successful when Charlie Ward plays point guard, a coach can issue a query for the identical circumstances, except with Derek Harper as point guard (this example also shows how AS can be used as a traditional query and report tool).

Perhaps the best opportunity to interpret a pattern is via video tape of the actual events. AS presents the video times for every interesting pattern. Coaches can then view just those segments of video tape. The video tape provides the complete context surrounding an interesting pattern and coupled with the perceptive abilities of a domain expert, interpretation and knowledge discovery is greatly facilitated. In the most recent version of AS, an entire game can be stored on a CD-ROM and the relevant segments can be isolated, accessed and viewed on a PC itself. Current trends to digital video should make this more cost effective and more prevalent in the near future.

References

Bhandari, I., Halliday, M., Tarver, E., Brown, D., Chaar, J., & Chillarege, R. (1993). "A case study of software process improvement during development." *IEEE Transactions on Software Engineering*, 19(12), 1157-1170

Bhandari, I. (1995). "Attribute Focusing: Data mining for the layman" *(Research Report RC 20136)*. IBM T.J. Watson Research Center.

Celko, J. & McDonald, J. (1995). "Don't warehouse dirty data." *Datamation*, v41, n19, p42(5).

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. (1996). "From data mining to knowledge discovery: An overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P.Smyth & R. Uthurusamy (Eds.)," *Advances in data mining and knowledge discovery* (pp 1-34). Cambridge, MA: MIT Press.

McMurray, S. (1995). "Basketball's new high-tech guru." *U.S. News and World Report*, December 11, 1995, pp 79-80.

Redman, T. (1992). *Data Quality: Management and Technology.* New York: Bantam Books.

Sterba, J. (1996). "Let's go to the digital videotape!" *Wall Street Journal*, March 22 1996, pg 7.

Ed Colet is currenty completing a Ph. D. In Cognitive Science at New York University and is currently also working at IBM's T.J. watson Research Center. His research interests focus on modeling the way people process quantitative information, and on data analysis and statistical methodology. He is the Senior Analyst in the Advanced Scout development group focusing on human factors, usability, underlying statistical methods, as well as broader issues that relate data mining software with information technology.

Inderpal Bhandari is a member of the research staff at the IBM T.J. Watson Research Center. He is currently directing data mining work of the Advanced Scout development group. Inderpal was educated at Carnegie Mellon University (Ph.D, 1990), the University of Massachusetts at Amherst

(M.S.) and the Birla Institute of Technology and Science, Pilani, India (B.Engg). His research interests are broad and he has published on topics relating to design automation, automated diagnostic methods, distributed algorithms and systems. His current research interests focus on data mining - specifically, automatic methods to simplify the process of discovery so that a layperson may discover knowledge from data and put that knowledge to practical use.

Jennifer Parker is presently at American Management Systems, Inc. developing an advanced imaging and workflow application for a leading insurance company in Hartford, CT. She is heavily involved in recruiting as well as team building efforts for the Insurance Technology Group of AMS, Inc. Jennifer received her computer science degree from Marist College in Poughkeepsie, NY and plans to further her education with an MBA concentrating in Information Systems.

Zachary Pines is a freshman at Yale University and a graduate of Briarcliff High School (Briarcliff Manor, NY). Zak has worked on the Advanced Scout project since May, 1995. His contributions include customizing Advanced Scout for use by NBA coaches, simplifying the user interface, and working with coaches to integrate the use of Advanced Scout into game preparation.

Rajiv Pratap is currently on assignment at the IBM T.J. Watson Research Center, Hawthorne from TISL, India (IBM, India). He was educated at the Indian Institute of Technology Kanpur,India. As the Technical Lead of the Advanced Scout team, Rajiv is responsible for the core design and development work. Additional responsibilities include marketing efforts and customer interaction. He is actively involved in extending data mining solutions to the Insurance, Cataloging and Retail industries. He is coordinating the design and development of customized solutions for these domains actively involved with customer and domain experts.

Krishnakumar Ramanujam is currently on assignment at the IBM T.J. Watson Research Center, Hawthorne from IBM, India. He obtained a Masters degree in Telecommunications at the Indian Institute of Technology, Bombay,India. He is involved in the design and development of the Advanced Scout program, and in extending the Attribute Focusing algorithm to other domains, notably the retail industry. He is also leading the design and development effort involving data mining and knowledge discovery on the Internet.