5C.6

DATA MINING AND KNOWLEDGE DISCOVERY OF DROUGHT IN NEBRASKA

Tsegaye Tadesse, Michael J. Hayes, and Don A. Wilhite
National Drought Mitigation Center, University of Nebraska - Lincoln, Lincoln, Nebraska

Sherri K. Harms
Department of Computer Science and Information Systems, University of Nebraska-Kearney, Kearney, Nebraska

1. INTRODUCTION

Drought impacts society in many ways. Studies show that economic, social, and environmental costs and losses associated with drought are increasing dramatically (Wilhite, 2000). According to the National Climatic Data Center (NCDC), the U.S. sustained 46 weather-related disasters in the period from 1980 to 1999, in which overall losses reached or exceeded $1 billion at the time of each event (Ross et al., 2000). Out of these disasters eight major droughts that occurred in the U.S. between 1980 and 1999 accounted for the largest percentage (43%) of weather-related monetary losses. The second largest percentage (30%) was due to hurricanes and tropical storms. In Nebraska, during the ten-year period from 1989 to 1998, the indemnity paid for drought losses totaled more than $92 million (USDA RMA, 1999). This implies that if droughts are frequent in the future, the losses could be much more than what has been observed in the 1980s and 1990s. Thus, since droughts are recurrent natural phenomena in Nebraska, it is obvious that proactive steps to address droughts are essential.

Studying past and present droughts in relation to climatological, oceanic, and atmospheric parameters could help mitigate future drought impacts on society by improving our understanding of the drought hazard. Large historical data sets are essential to identify relationships between different climatic parameters and to distinguish patterns that may be used to predict drought. In light of this, it is essential to have an efficient way to extract information from large databases and to deliver relevant and actionable information for drought mitigation. One of the recently developed techniques relevant for such purposes is "Data Mining."

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships of physical variables in different data sets (Two Crows, 1999). This technique is used in multidisciplinary fields bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases (Cabena et al., 1998). The method is used for commercial applications by many companies for designing strategic benefits to increase profitability (Groth, 1998).

* Corresponding author address: Tsegaye Tadesse, National Drought Mitigation Center, Univ. of Nebraska - Lincoln, NE 68583; e-mail: ttadesse2@unl.edu.

The aim of this study is to develop time-series algorithms that can identify the relationships of the oceanic parameters and drought so that decision makers would use data-driven knowledge based decisions to reduce the impacts of drought.

2. TIME-SERIES DATA MINING ALGORITHMS

In this study, two data mining algorithms are developed to identify the relationships of drought and oceanic indices. These algorithms are discussed below.

2.1 Representative Episodes Association Rules (REAR) algorithm

Representative Episodes Association Rules (REAR) algorithm (Harms et al., 2001) converts the time-series data into discrete representations and generates association rules. The algorithm counts occurrences of the drought episodes that occur together with other oceanic parameters within the sliding window. Then the algorithm sets aside the frequent episodes that occur at more than the predetermined minimum frequency (Harms et al., 2002). The rules are generated using the antecedent (*A*) and consequence (*B*) constraints to keep track of the events. This identifies the *A* and *B* of the association rules in the form of: *if A then B* with minimum confidence. The rules that are generated using REAR can be all of the rules or representative of all rules, depending on the user requirements.

Representative association rules provide a smaller set of rules that still shows the 'interesting' ones (Harms, 2001). The number of rules that are generated depends on the minimum frequency, window width, and minimum confidence values. In selecting these parameters, one may have to consider the advantages and drawbacks of the parameters on the outputs (i.e., rules that are generated). For example, if a wider window width is selected, more relationships may be found, but the analysis and interpretation of the rules may be difficult. If a smaller frequency is chosen, there could be more rules with high confidence, but since the episodes are not frequent, they may be less meaningful.

2.2 The Minimal Occurrences With Constraints and Time Lags (MOWCATL) algorithm

The Minimal Occurrences With Constraints and Time Lags (MOWCATL) algorithm was developed to find relationships between sequences in multiple data

sets. Similar to the REAR approach, MOWCATL uses antecedent and consequence constraints in generating rules to keep track of the events. In sequential data mining, MOWCATL uses separate antecedent and consequent inclusion constraints, along with separate antecedent and consequent maximum window widths, to specify the antecedent and consequent patterns that are separated by a time lag. This approach is based on the concepts of association rules combined with frequent episodes, time lags, and event constraints (Harms et al., 2002). The advantage of this algorithm over the REAR method is that it adds time lags.

The MOWCATL has three window parameters: the maximum window width of the antecedent (win_a), the maximum window width of the consequent (win_c), and the time lag. Using these parameters, the algorithm generates episodal rules where the antecedent episode occurs within a given maximum window width, the consequent episode occurs within a given maximum window width, and the start of the consequent follows the start of the antecedent within a given maximum time lag. The generated rules are such that if *A* and *B* occur within 2 months, then within 3 months they will be followed by *C* and *D* occurring together within 2 months.

The MOWCATL approach is well suited for sequential data mining problems that have groupings of events that occur close together, but occur relatively infrequently over the entire data set. It is also well suited for problems that have periodic occurrences when the signature of one or more sequence is present in other sequences, even when the multiple sequences are not globally correlated. The analysis techniques developed in this study facilitate the evaluation of the temporal associations between episodes of events and the incorporation of this knowledge into decision support systems. This new approach complements the existing approaches to address drought research problems (Harms et al., 2002).

3. DROUGHT EPISODE IDENTIFICATION

In this study data mining techniques are introduced to help in understanding and identifying complex relationships of atmospheric and environmental variables causing droughts. Using data mining tools, one can identify 'local' patterns better than the traditional time-series analysis techniques that largely focus on global models such as statistical correlations (Das, 1998). The infrequent and complex nature of drought requires alternative analysis techniques that emphasize the discovery of local patterns of climate and oceanic data. For example, one may consider the occurrence of drought and its association with climatic and oceanic parameters instead of all precipitation patterns that include wet periods as well. In other words, since drought monitoring is particularly concerned with drought episodes, the data-mining algorithm is needed to discover the associations of drought with oceanic and atmospheric conditions causing drought. This algorithm should identify the drought episodes without the distractions of other 'non-interesting' episodes that include normal precipitation and wet episodes within the time series (Harms et al., 2001).

The ocean-atmosphere relationship is considered to have an impact on local droughts in studying the associations between drought and the oceanic and climatic indices. Since global oceanic parameters such as sea surface temperatures are changing relatively slower than the surface climatic parameters such as precipitation and temperature, one can better understand the trend of the oceanic parameters as compared to the highly variable and infrequent drought episodes. Thus, the oceanic parameters are selected to be used as antecedents and droughts as consequents in finding their associations in this study.

4. RESULTS OF CLAY CENTER, NEBRASKA

Association rules were generated using the two developed data mining algorithms (i.e., REAR and MOWCATL) for five selected stations and state-average data of Nebraska for the period of fifty years, 1950 to 1999. In this paper, association rules that were generated and analyzed for Clay Center, Nebraska is used to demonstrate the results.

The droughts are defined based on the values of the Standardized Precipitation Index (SPI) and Palmer Drought Severity Index (PDSI) values. The other oceanic and climatic parameters used for the same period of study (1950-99) are the Southern Oscillation Index (SOI), the Multivariate ENSO Index (MEI), the Pacific/North American (PNA), the Pacific Decadal Oscillation (PDO), and the North Atlantic Oscillation (NAO).

The rule generated for Clay Center using the REAR algorithm for parallel episodes of one-month window width include: if the SOI value was between 1 and 1.5, MEI value was less than -1.5, and the PDO was less than -2, then the PDSI was extremely dry with 83% minimum confidence.

Using MOWCATL, with antecedent and consequent windows of size of 1 month and a maximum one-month time lag between the start of the antecedent oceanic parameters and the start of the consequent drought episodes, the rules generated for Clay Center, Nebraska, include: if the SOI was between 1 and 1.5, the MEI was less than –1.5, and the PDO was less than –2, then the nine-month SPI was severely dry, the twelve-month SPI was severely dry, and the PDSI was extremely dry with more than 83% confidence.

The results showed that most occurrences of drought based on the SPI and PDSI categories are associated with the SOI, MEI, and PDO with different combinations and confidence factors. The combinations of negative MEI values (La Niña), positive values of SOI (La Niña), and negative PDO values implied occurrences of droughts.

5. CONCLUSION

The rules generated using both algorithms indicate that there are strong associations between the oceanic indices and droughts occurrence over the selected five

stations and state-averaged data of Nebraska. Most of the rules that are generated using both data mining algorithms indicate that there are strong relationships between the drought episodes in Nebraska and the SOI, MEI, and PDO indices.

This study has also identified three major advantages of data mining as compared to the previous traditional statistical methods: (i) instead of global correlation of the climatic and oceanic data, target episodes such as droughts can be specified separately from normal and wet conditions, (ii) data mining algorithms give flexibility in time-series analyses, allowing the discovery of relationships of the parameters with time lags using sliding windows, and (iii) the algorithms allow the analysis of large amounts of data and complicated computations to be executed within a reasonable period of time.

Because of its flexibility in time, data mining algorithms that include the time lag factor (e.g. MOWCATL) identify a better association of the oceanic and climatic parameters to predict drought. Since the generated rules indicate the occurrence of drought given certain oceanic parameters, the data mining algorithms that identify drought episodes and associate the parameters with a time lag are robust tools in drought monitoring.

The association rules provide the information that the oceanic parameters could be used as a precursor of drought. However, it is important to note that the newly developed data mining techniques of identifying drought episodes based on drought's associations with oceanic and climatic parameters are intended to provide additional drought monitoring tools to complement other common techniques already in use.

REFERENCES:

Cabena, P.H., R. Stadler, J. Verhees, and A. Zanasi, 1998. Discovering data mining: from concept to implementation. IBM, New Jersey, pp.195.

Das, G., K.I. Lin, and H. Mannila, 1998. Rule discovery from time-series. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pp. 16-22, New York, NY.

Groth, R., 1998. Data Mining: a hands-on approach for business professionals. Prentice Hall Inc., New Jersey, pp. 264.

Harms, S.K., J. Deogun, J. Saquer, and T. Tadesse, 2001. Discovering Representative Episodal Association Rules from Event Eequences Using Frequent Closed Episode Sets and Event Constraints. In the Proceedings 2001 IEEE International Conference on Data Mining (ICDM '01), San Jose, California.

Harms, S.K., J. Deogun, and T. Tadesse, 2002. Discovering Sequential Rules with Constraints and Time Lags in Multiple Sequences. In proceedings of 2002 International Symposium on Methodologies for Intelligent Systems, Lyon, France.

Ross, T., and N. Lott, 2000. A climatology of recent extreme weather and climatic events, US Dept. of Commerce, NOAA/NESDIS, National Climatic Data Center (NCDC), Technical Report 2000-02, pp.17. Available on line at **http://lwf.ncdc.noaa.gov/oa/reports/billionz.html**, (April 15, 2001).

Two Crows Corporation, 1999. Introduction to data mining and knowledge discovery, third ed., Postmac, MD. Available at: **www.twocrows.com**, (April 29, 2000).

USDA RMA, 1999. USDA National Risk Management Agency database. Risk management office, Billings, MT.

Wilhite, D.A., 2000. Drought as a natural Hazard: concepts and definitions. Drought: A Global Assessment, Ed. D.A. Wilhite, Routledge Hazards and Disaster Series, Vol. I, 3-18.