Time-Series Data Mining in a Geospatial Decision Support System[*]

Dan Li, Sherri Harms, Steve Goddard, William Waltman, Jitender Deogun
Department of Computer Science and Engineering
University of Nebraska-Lincoln, Lincoln NE 68588-0115

Abstract

This paper presents an overview of the motivation for, and the use of time-series data mining in, a Geospatial Decision Support System (GDSS). Our approach is based on a combination of time-series data mining algorithms and spatial interpolation techniques. The initial focus of the system is to facilitate drought risk management. We develop two association rule mining algorithms and two interpolation methods, which help drought experts predict local weather conditions or potential yield impact based on the global weather patterns.

Keywords: Geospatial Decision Support System, Data Mining, Interpolation.

1. Introduction

Drought is a natural process of Great Plains landscapes and results in significant economic, social, and environmental impacts. Historically, more emphasis has been placed on the response component of drought management, with little or no attention to mitigation, preparedness, and prediction or monitoring (Wilhite 2001). Thus, through the National Science Foundation (NSF) Digital Government program, the USDA RMA is working with the University of Nebraska–Lincoln Computer Science and Engineering (CSE) Department, National Drought Mitigation Center (NDMC), and High Plains Regional Climate Center (HPRCC) to develop a Geospatial Decision Support System (GDSS) to improve the quality and accessibility of temperature and precipitation data for drought assessment and drought risk management. Figure 1 shows two drought assessment maps. The drought map of Nebraska can be generated in real-time produced by our GDSS system for any specified time interval from the project's home page: http://nadss.unl.edu/.

A common question in risk analysis is "How are events related in time?" In a risk management application where a time-series is a factor, it is important to study the relationships of the parameters that occur together. Data mining algorithms have the potential to identify these relationships. Predicting events and identifying sequential rules that are inherent in the data help domain experts learn from past data and make informed decisions for the future. For example, decision-makers are interested in discovering associations between the periodical occurrence of El Niño and the periodical occurrence of natural hazards. Data mining techniques can help us build abstract models to represent the reality and to support risk management and mitigation of natural hazards. In the rest of this paper, we demonstrate the integration of spatio-temporal knowledge discovery techniques in the GDSS using a combination of data mining methods applied to geospatial time-series data.

[*]This research was supported in part by NSF Digital Government Grant No. EIA-0091530 and NSF EPSCOR, Grant No. EPS-0091900.

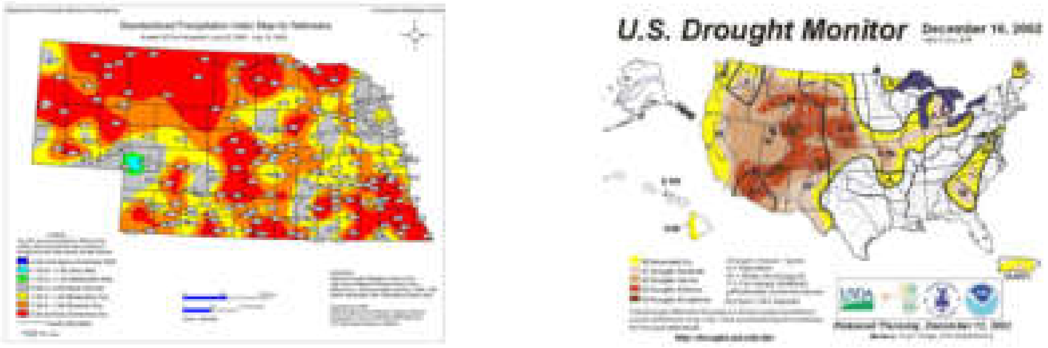

Figure 1: Drought assessment maps produced by the Computer Science and Engineering Department and the National Drought Mitigation Center at UNL for the USDA and NOAA. The goal of this project is to increase the resolution and map potential impact of hazards.

2. Data Mining Algorithms

The basis of all relationship detection by data mining is essentially association rule mining. Association rules are relations between variables of the form $X \Rightarrow Y$. The problem was first defined in the context of the market basket data to identify customer buying habits (Agrawal et al. 1993). For example, it is of interest to a supermarket manager to find that 80% of the customers who buy bagels also buy cream cheese and 5% of all customers buy both bagels and cream cheese. Here the association rule is $bagels \Rightarrow cream - cheese$, 80% is the *confidence* of the rule and 5% is its *support*. For drought risk management, we want to discover similar association rules which reflect the relationships between environmental variables and drought events. In such association rules, the variable X is an antecedent episode (e.g. Multivariate ENSO Index–MEI), and Y is a consequent episode (e.g. Standardized Precipitation Index–SPI, or corn yield). An episode is an event sequence. Figure 2 shows two event sequences, MEI and corn yield in Nebraska, from 1950 to 2000. Association rules capture the associations between the variation trend shared by these sequences.

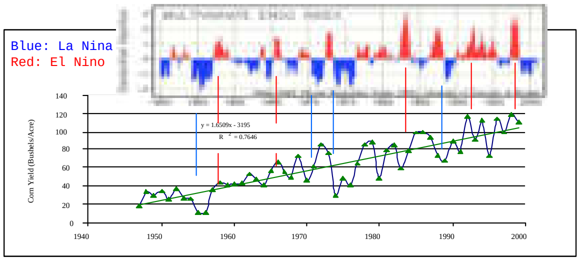

Figure 2: Association between MEI and corn yield through time in Nebraska.

We have developed two data mining algorithms, *REAR* (Harms et al. 2001) and *MOWCATL* (Harms et al. 2002), to generate association rules. The algorithms demonstrate the importance and potential use of data mining techniques in monitoring drought using the oceanic and atmospheric indices. These algorithms have been effectively employed in the drought assessment problem to find relationships between global climatic episodes and local drought conditions. The relationships can then be used to predict target drought episodes and potential yield impact based on oceanic indices such as MEI.

3. Interpolation Techniques

To facilitate drought risk management, we collect climatic data from a variety of sources at weather stations. However, it is impossible to collect climate data everywhere due to cost and physical considerations. To make our spatio-temporal data mining more interesting and meaningful to a broader geographical area, we extend our work to discover association rules for areas not covered by existing weather stations. Since spatial interpolation has the potential to find a function that will predict data values at unsampled points given a set of spatial data at sampled points, we designed and implemented two interpolation approaches (Li et al. 2003), pre-interpolation and post-interpolation, to best facilitate drought risk management.

Figure 3 shows the differences between pre-interpolation and post-interpolation methods. In *pre-interpolation* approach, we apply appropriate interpolation methods (i.e., Inverse Distance Weighting (IDW) and Kriging) to obtain datasets for query points, then, working on interpolated data, association rules are discovered for these points. To obtain greater accuracy, we modify the weight function in the basic IDW method to accommodate some geographic and climatic features. In the *post-interpolation* approach, we first discover association rules for sample points, then propose an interpolation algorithm to discover association rules for query points.

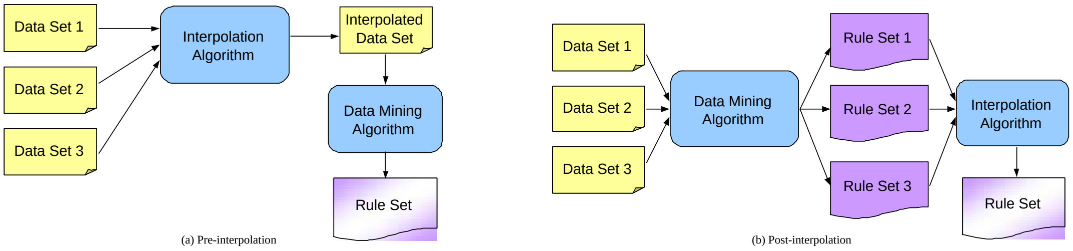

Figure 3: Two interpolation methods.

To evaluate the performance of our interpolation methods, we borrow two widely used quality metrics, precision and recall. *Precision* is defined as the percentage of actual rules correctly discovered among all discovered rules by a certain interpolation algorithm. *Recall* is defined as the percentage of actual rules discovered by an interpolation method to the number of actual rules discovered with sample datasets. Our experiments show that the post-interpolation method gives higher precision while the pre-interpolation method provides higher recall. Among the three pre-interpolation methods evaluated, i.e. basic IDW, modified IDW, and Kriging, the Kriging method outperforms the others, and the modified IDW method presents better results than the basic IDW method.

4. A Conceptual Model for Intelligent Report Generation

Although association rule mining provides a way of discovering hidden patterns in large volumes of data, association rules themselves are too abstruse to be understood, especially for a person who is not a domain expert. For example, given an association rule r : MEI extremely dry, PDO extremely dry $\Rightarrow$ PDSI severely dry ($confidence = 65\%$). From this rule, we can not see to what degree a drought will occur if both the MEI and the PDI are in the extremely dry condition. To address this problem, we propose an automated interpreter to translate association rules into plain English reports. The above rule can be explained as: if both MEI and PDO are extremely dry, then with 65% possibility, PDSI will be severely dry. This rule can be interpreted as: because it is going to be severely dry, one should plan to plant a crop that is suitable for dry weather. Figure 4 shows a conceptual model for intelligent report generation that includes a proposed interpreter. The development of the intelligent interpreter will help non-expert users in effective decision making.

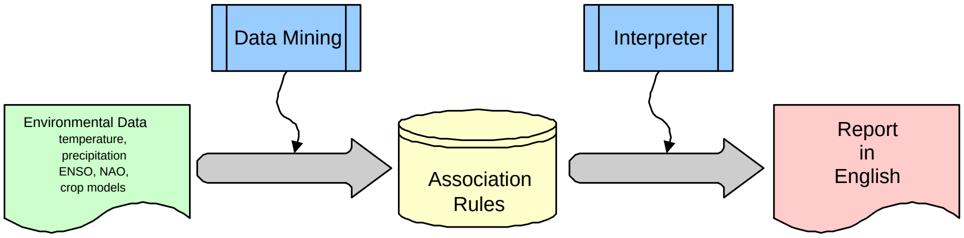

Figure 4: A conceptual model for intelligent report generation

5. Conclusion

This paper presents the application of time-series data mining techniques in a Geospatial Decision Support System (GDSS) for drought risk management. We developed two association rule mining algorithms, REAR and MOWCATL, which find relationships between global climatic episodes and local drought conditions. For the sites that do not have sampled data, we investigated two spatial interpolation approaches, pre-interpolation and post-interpolation, which facilitate the process of discovering association rules. The discovered rules help drought experts predict local weather conditions or potential yield impact based on the global weather patterns.

References

R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD 1993 International Conference on Management of Data [SIGMOD 93]*, pages 207–216, Washington D.C., 1993.

S. Harms, J. Deogun, J. Saquer, and T. Tadesse. Discovering representative episodal association rules from event sequences using frequent closed episode sets and event constraints. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 603–606, San Jose, California, USA, November 29 - December 2 2001.

S. Harms, J. Deogun, and T. Tadesse. Discovering sequential association rules with constraints and time lags in multiple sequences. In *Proceedings of the 2002 International Symposium on Methodologies for Intelligent Systems (ISMIS '02)*, pages 432–442, Lyon, France, June 2002.

D. Li, J. Deogun, and S. Harms. Interpolation techniques for geo-spatial association rule mining. In *Proceedings of the 9th. International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, Chongqing, China, May 26-29 2003. (to appear).

D. A. Wilhite. Moving beyond crisis management. *Forum for Applied Research and Public Policy*, 16(1):20–28, 2001.