Mining Conditions in Rapid Intensifications of Tropical Cyclones ---A successful example of scientific data mining---

Jiang Tang [1], Ruixin Yang [1] *, Daniel Barbara [2], Menas Kafatos [1]
[1] Center for Earth Observing and Space Research, College of Science, George Mason University
[2] Department of Information and Software Engineering, George Mason University
{jtang, ryang, dbarbara, mkafatos}@gmu.edu

Abstract

Rapid intensification (RI) of tropical cyclones (TC) is a major error source in TC intensity forecasting. Unexpectedly strong storms can cause substantial damages and losses. In order to improve RI probability estimation, the association rule is used in this study in order to mine candidate sets of factors which have strong interactions with rapidly intensifying TCs. Our mining results identified a reduced predictor set with fewer factors identified in previous studies but improved RI probabilities. This is a real, successful example of scientific data mining.

1. Introduction

Tropical cyclones are a type of often costly natural hazard. Unanticipated landfalls and storm intensities are the instrumental causes of substantial damages and losses. The situation can become even worse for rapidly intensifying tropical cyclones. A tropical cyclone (TC) is said to undergo rapid intensification (RI) if its intensity has increased at least 15.4 m/s (30 knots) over a 24-hour period [1]. Predicting the potential probability of rapid intensification of a tropical cyclone remains an urgent demand of the coastal populations, governments, insurance companies and scientists alike.

The rapid intensification of a TC is likely influenced by the warm sea surface temperature, the TC's inner-core processes, and the environmental flow interactions such as weak vertical shear and the enhanced relative angular momentum of an upper-level trough, etc. [1,2,3,4,5,6,7,8]. More recently, Kaplan and DeMaria [1] (hereafter KD03) examined the large-scale characteristics of rapidly intensifying Atlantic tropical cyclones from 1989 to 2000. They developed a scheme to estimate the RI probability by combining five persistent and synoptic conditions: the previous 12-hr intensity change (DVMX) >= 4.6m/s, the vertical shear (SHR) <= 4.9m/s, the sea surface temperature (SST) >= 28.4^{o}C, the difference between the current intensity and the maximum potential intensity (POT) >= 47.6m/s, and the low-level relative humidity (RHLO) >= 69.7%. However it is subjective and labor-intensive to pick these five conditions from a total of 11 conditions. Therefore, we sought to develop an objective method that can reveal statistically-sound condition combinations for estimating RI probability.

Association rule algorithms (AR), originally developed by Agrawal *et al.* [9, 10], are considered as a new generation of multi-correlation discovery algorithms. In this paper we leverage the association rule technique to mine combinations of conditions for what appeared in rapidly intensifying tropical cyclones. Data mining is a promising tool for exploratory scientific data analysis. Unfortunately, the results derived from the mining outputs are usually known phenomena [e.g., 11]. In other words, scientific data mining results often confirmed what scientists already know without giving much new knowledge. In this paper, we show an example which reveals a larger RI probability with fewer factors identified than by traditional statistical analysis.

2. Data

The datasets used for this study are the NHC HURDAT file [12] and the SHIPS 1989-2000 database [5]. The methodology for merging the two datasets is identical to what was described in KD03 except that we ignore the location-based sample selection and

* Corresponding author.

exclude the non-developing tropical depressions. After the merge, the reconstructed dataset contains a total of 3306 observations from the 1989-2000 period, which were from 135 distinct Atlantic TCs (0 tropical depressions, 54 tropical storms, and 81 hurricanes). Abiding by the RI definition proposed in KD03, that is, at least 30 knots of intensity increase within 24 hours, the 3306 observations comprise 169 RI cases and 3137 non-RI cases, thus the sample mean RI probability is 5.11%.

Eleven parameters were used in KD03 because of their statistically significant differences between RI cases and non-RI cases. These parameters are the previous 12-hr intensity change (DVMX), vertical shear (SHR), sea surface temperature (SST), the difference between the current intensity and the maximum potential intensity (POT), the 850-700hPa relative humidity (RHLO), latitude of current TC location (LAT), longitude of current TC location (LON), the zonal component of storm motion (USTM), the zonal component of upper-level wind (U200), the upper-level relative angular momentum flux (REFC), and the environmental steering pressure level (SLYR). A detailed description of the 11 variables in HURDAT and SHIPS datasets can be found in KD03.

3. Methods

There are two high level approaches for conducting scientific data mining. One approach is to mine the scientific data, usually multidimensional spatio-temporal arrays directly following the data models used in scientific domains [e.g., 13]. The other approach is to convert the scientific data into transaction data and then use the classical data mining tools such as association rules to reveal hidden relationships between different physical parameters. In this study, we follow the second approach by employing the association rule data mining algorithm to find frequent condition sets associated with RI of TCs.

The version of the association rule algorithm we used in this study is the Apriori algorithm implemented by Borgelt [14]. The support value in this implementation is defined as probability of antecedents only instead of probability of antecedents and the consequent, and this definition is the same as those used in traditional analysis such as that used in KD03. Since no previous experience exists on this specific data mining task for TCs undergoing RI, no specific control parameters (predefined support, confidence, and lift) are chosen. Instead, results are compared to the results of KD03, and the rules giving significant results are discussed.

Before feeding the data into the mining algorithm, preprocessing is necessary for converting the geophysical values into a transaction-like data set. In the preprocessing step, each of the 11 continuous persistent and synoptic attributes are discretized into "Low" and "High" ranges using the same threshold values as KD03 for the consistency consideration. Also, abiding by the RI definition in KD03, the target attribute (the future 24-hr intensity change, FD24) is transformed into class RI and class non-RI.

After preprocessing, the Apriori method is applied to find all closed large condition sets among these attributes. Therefore a closed frequent condition set containing the condition "FD24=RI" and other persistent and synoptic attributes indicates an association among these attributes and the future rapid intensification. The process of finding a set of predictors which have improved RI probabilities is accomplished by pruning the association rules.

4. Results and discussion

Kaplan and DeMaria [1] found that higher RI probabilities can be obtained when a combination of conditions is satisfied and identified the highest RI probabilities occurred when five conditions (high DVMX, low SHR, high SST, high POT, and high RHLO) are satisfied together. They claimed that about 2% of the total samples satisfy the condition combination and among them 41% underwent RI.

The above cited result can be easily translated into an association rule as "*RI <= SHR=L, DVMX=H, SST=H, POT=H, RHLO=H*" with support of the antecedent at 2% and confidence at 41%. Indeed, the exact associate rule with support of the antecedent at 0.7%, confidence at 43.5% and lift at 850.5% is found by using the association rule technique. The lower support is likely due to the larger portion of non-RI samples in our dataset. However the higher confidence is likely due to the omission of the non-developing depressions. We suspect that the non-developing depressions may contain a higher proportion of cases satisfying the five conditions but without undergoing RI, which resulted in the confidence of 43.5% in our study, higher than the reported 41% from KD03.

Since our goal is to mine condition sets with an improved RI probability, we use the pruning procedure to the frequent sets. During the pruning process, the original rule with the five KD03 constraints satisfied, "*RI⇐ SHR=L, DVMX=H, SST=H, POT=H, RHLO=H (supp=0.7%, conf=43.5%, lift=850.5%),*" is removed

as a redundant rule from the rule set. A more interesting rule with three constraints satisfied, *"RI⇐ SHR=L, DVMX=H, RHLO=H (supp=1.3%, conf=47.6%, lift=931.5%),"* remains. This means that this rule with all 5 constraints satisfied in KD03 is not the best and a rule with fewer constraints fits more RI cases. Although the remaining rule has a shorter list of antecedents, it has a higher confidence than the original rule, which indicates that the 3 constraints, SHR=L, DVMX=H, and RHLO=H, can explain more general RI cases than with all of the 5 constraints together. Therefore, increasing the number of constraints will not improve the accuracy of the determination of RI cases. Meanwhile, without a proper guidance, it is difficult to select RI constraints in the RI probability estimate to achieve the optimal results. The association rule mining with the pruning process provides a systematic implementation to reach this goal.

To compare the results with those in KD03, the composite RI probability defined in KD03 is computed. KD03 defined the composite RI probability as a function of the total number of the given conditions that were satisfied (as Figure 9 in KD03). That is, given a set of conditions with length N, (N+1) probabilities are computed as follows:

- When n=0, the RI probability is defined as the conditional probability of RI given that none of the N conditions was satisfied;
- When n=1, the RI probability is defined as the conditional probability of RI given that only 1 of the N conditions were satisfied;
- When n=2, the RI probability is defined as the conditional probability of RI given that only 2 of the N conditions were satisfied;
- Similar calculation continues until n=N.

Figure 1 displays the RI probabilities with five conditions identified in KD03 and the three conditions mined out in this work. It is evident that for both cases, the RI probabilities are higher than the sample mean probability when at least one condition in either case is satisfied. Moreover, both curves show the monotonic increasing trend with the number of predicting conditions. And there is no significant difference between the RI probabilities with at least one constraint satisfied for the cases of the five identified constraints and of the three reduced constraints.

However, when at least two constraints are satisfied, the probabilities with three identified constraints are significantly higher than the corresponding probabilities in the five constraint case. The most significant result of this work is that when all three identified constraints are satisfied, the RI probability is higher than the probability with all five satisfied constraints identified in KD03. The RI probability with all five KD03 conditions satisfied is 43.5%, and the corresponding number for only the three conditions mined out rises to 47.6%. This result is significant not only to the study of rapid intensification of hurricanes but also for the data mining procedures in identifying meaningful scientific results. The results successfully demonstrate that the association rule data mining technique can be used as an exploration method to generate hypotheses, and a statistical analysis should be performed as confirmation of the hypotheses, as generally expected for data mining applications [15]. The data mining results also shed light for potential improvement of TC intensity forecasting.

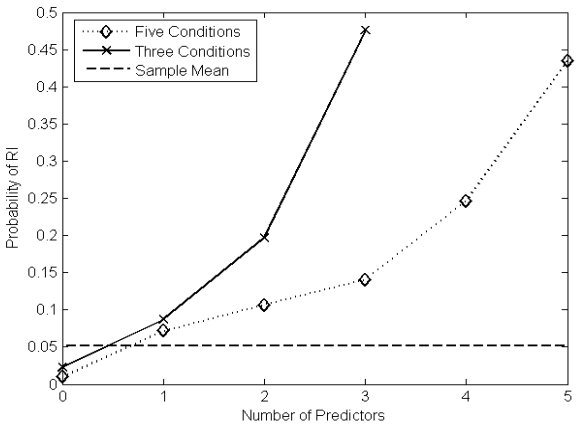

Figure 1. The composite RI probabilities based on the condition set in KD03 {DVMX=H, SST=H, POT=H, RHLO=H, SHR=L} and the three-mined-condition set {DVMX=H, RHLO=H, SHR=L}. The sample mean RI probability is also shown for reference.

5. Conclusion

Researchers have developed various methods to predict the intensity of a tropical cyclone [5]. Many of those methods suffer degradation of performance for rapidly intensifying tropical cyclones [5]. In this study, we applied the association rule to look for combinations of persistent and synoptic conditions which may provide improved RI probability estimates.

Compared to statistical analysis, the technique of association rules can explore associations among multiple conditions without extra effort because it examines all possible combinations of frequent condition sets automatically. It provides an as complete as possible picture of the dataset to scientists so that the connections among multiple conditions will

not be overlooked by a theory-driven analysis approach. Compared to the statistical analysis of KD03, the data mining technique used in this work not only identified the predictors giving an improved RI probability but also obtained this result with fewer predictors through a pruning process of association rules. This work demonstrates that the association rule data mining technique can be used as an exploration method to generate hypotheses, and a statistical analysis should be performed as confirmation of the hypotheses, as generally expected for data mining applications.

This work shed light on the physical conditions favoring RI processes of TCs. To have a more comprehensive screening of those conditions with data mining algorithm is a challenge. Actually, the parameter values from SHIPS data include those from diverse data sources such as observations and numerical weather forecasting. Converting these geophysical values, usually as multi-dimensional arrays, into transaction type data in large volume is not a trivial task. Better cyber-infrastructure with intelligent archiving features such as server-side data manipulation and data aggregation may significantly reduce the time-demanding preprocessing load. Before we have such a cyber-infrastructure with the right data, data mining can only be efficiently used for limited data sources which are well defined and can be easily handled.

6. Acknowledgements

The authors would like to thank Dr. Mark DeMaria for providing the 2003 SHIPS data and Dr. Christian Borgelt for making his implementation of the Apriori association rule algorithm available. The authors also benefit from discussions with Drs. Mark DeMaria, Donglian Sun, and Liguang Wu. This work was partially supported by NASA Grant NNX06AF30G.

7. References

[1] J. Kaplan, and M. DeMaria, "Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin," *Weather and Forecasting*, 2003, vol.18, pp 1093-1108.

[2] H. E. Willoughby, J. A. Clos, and M. G. Shoreibah, "Concentric eyewalls, secondary wind maxima, and the evolution of the hurricane vortex," *Journal of the Atmospheric Science*, 1982, vol.39, pp 395–411.

[3] W. M. Gray, "Global view of the origin of tropical disturbances and storms," *Monthly Weather Review*, 1968, vol.96, pp 669–700.

[4] R. T. Merrill, "Environmental influences on hurricane intensification," *Journal of the Atmospheric Science*, 1988, vol. 45, pp 1678–1687.

[5] M. DeMaria, and J. Kaplan, "A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin," *Weather and Forecasting*, 1994, vol. 9, pp. 209-220.

[6] M. DeMaria, "The effect of vertical shear on tropical cyclone intensity change." *Journal of the Atmospheric Science*, 1996, vol.53, pp 2076-2087.

[7] DeMaria, M., J.-J Baik, and J. Kaplan, "Upper-level eddy angular momentum flux and tropical cyclone intensity change," *Journal of the Atmospheric Science*, 1993, vol.50, pp 1133-1147.

[8] C. R. Holliday, and A. H. Thompson, "Climatological characteristics of rapidly intensifying typhoons," *Monthly. Weather Review*, 1979, vol. 107, pp 1022–1034.

[9] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD 1993*, Washington D.C., May 1993, pp. 207-216.

[10] Han, J., and M. Kamber, "*Data Mining: Concepts and Techniques*," Morgan Kaufmann, San Francisco, 2001, 550 pp.

[11] M. Steinbach, P. Tan, V. Kumar, S. Klooster, and C. Potter, "Discovery of climate indices using clustering." In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington, D.C., August 24-27, 2003, Pages: 446-455.

[12] B.R. Jarvinen, C.J. Neumann, and M.A.S. Davis, "A tropical cyclone data tape for the North Atlantic basin, 1886-1983: Contents, limitations, and uses," *NOAA Technical Memorandum NWS NHC 22*, 1984.

[13] M. Steinbach, P.-N. Tan, V. Kumar, C. Potter, and S. Klooster, "Data mining for the discovery of ocean climate indices." In *Mining Scientific Datasets Workshop, 2nd Annual SIAM International Conference on Data Mining*, April 2002.

[14] Borgelt, C., *Apriori - Association Rule Induction / Frequent Item Set Mining*, http://www.borgelt.net/apriori.html, last access on August 24, 2007.

[15] Hand, D., H. Mannila, and P. Smyth, *Principles of Data Mining*. Cambridge, M.I.T Press, 2001