The 19th Conference on Probability and Statistics, American Meteorological Society, *New Orleans, LA, 20-24 January 2008*

7.3 MINING "OPTIMAL" CONDITIONS FOR RAPID INTENSIFICATIONS OF TROPICAL CYCLONES

Ruixin Yang*, Jiang Tang, Menas Kafatos
Center for Earth Observing and Space Research
Department of Earth Systems and GeoInformation Sciences
College of Science
George Mason University
Fairfax, VA 22030-4444

ABSTRACT

Rapidly intensifying (RI) tropical cyclones (TC) are the major error sources in TC intensity forecasting. In order to improve the estimates of RI probability, association rules as a data mining technique are used here to facilitate the process of looking for candidate sets of conditions which have strong interactions with rapidly intensifying TCs. Compared to the relation analysis method, the technique of association rules can easily explore associations among multiple conditions. Moreover, our mining results identified a reduced predictor set with fewer factors and improved probabilities of RI estimates compared to the results based on relation analysis. For a given number of constraints affecting the RI process, the data mining technique can identify the combination of the factors which give the largest RI probabilities. In this paper, we will present the main findings based on the data for SHIPS (Statistical Hurricane Intensity Prediction Scheme), an operational statistical-dynamical hurricane intensity forecasting mode.

1. INTRODUCTION

Tropical cyclones are one of the most costly natural disasters. Although TC track forecasting is of relatively high skill, intensity forecasting is still a challenge because there are many factors controlling TC intensity changes (DeMaria et al. 2007). The intensity forecasting can become even more difficult for rapidly intensifying tropical cyclones. A 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 (Kaplan and DeMaria 2003). The rapid intensification of a TC is likely influenced by warm sea surface temperature, the TC's inner-core processes, and environmental flow interactions such as weak vertical shear and the enhanced relative angular momentum of an upper-level trough, etc. (Gray 1968; Holiday and Thompson 1979; Willoughby et al. 1982; Merrill 1988; DeMaria et al. 1993; DeMaria and Kaplan 1994; DeMaria 1996; Lee and Bell 2007).

More recently, Kaplan and DeMaria (2003) (hereafter KD03) examined the large-scale characteristics of rapidly intensifying Atlantic tropical cyclones from 1989 to 2000. They developed a scheme to identify a high RI probability by combining five persistent and synoptic conditions: the previous 12-hr intensity change (PD12) >= 4.6m/s, the vertical shear (SHRD) <= 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 22 conditions based on 11 variables. Therefore, we sought to use an objective data mining method that can systematically reveal statistically-sound condition combinations for detecting higher RI probabilities.

One of such methods is the Association Rule algorithms, originally developed by Agrawal et al. (1993), which are considered as a new generation of multi-correlation discovery algorithms. In this paper we leverage the association rules technique to mine combinations of conditions for what appeared

* Corresponding author address: Ruixin Yang, MS 6C3, College of Science, George Mason University, Fairfax, VA 22030; e-mail: ryang@gmu.edu.

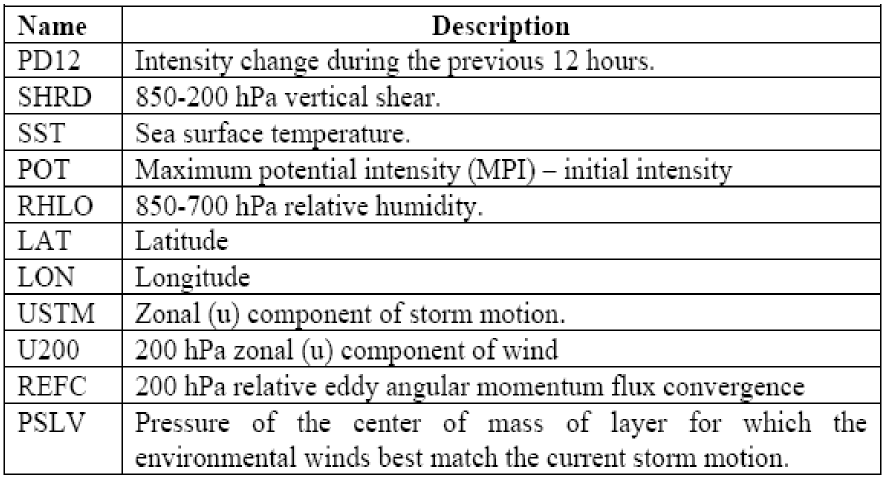
in past rapidly intensifying tropical cyclones. The goal of this research is to reveal the "optimal" condition for RI process of TCs.

2. DATA

The datasets used for this study are the NHC HURDAT file (Jarvinen et al. 1984) and the SHIPS 1989-2000 database (DeMaria et al. 2005). 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 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). Among the 3306 cases, only 169 of them represent RI cases with a sample mean RI probability being 5.11%.

Eleven parameters were used in KD03 because of their statistically significant differences between RI cases and non-RI cases. These parameters are listed in Table 1.

. Table 1. The 11 statistically significant predictors based on KD03 (Table 4). The predictors DVMX, SHR, and SLYR in KD03 are renamed as PD12, SHRD, and PSLV here.



Name	Description
PD12	Intensity change during the previous 12 hours.
SHRD	850-200 hPa vertical shear.
SST	Sea surface temperature.
POT	Maximum potential intensity (MPI) – initial intensity
RHLO	850-700 hPa relative humidity.
LAT	Latitude
LON	Longitude
USTM	Zonal (u) component of storm motion.
U200	200 hPa zonal (u) component of wind
REFC	200 hPa relative eddy angular momentum flux convergence
PSLV	Pressure of the center of mass of layer for which the environmental winds best match the current storm motion.

3. METHODS

To discover "multiple-to-one" associations among a large number of factors favoring rapidly intensifying TCs, we will "mine" the above cleaned data by using an association rule algorithm. Association rule induction (Agrawal et al. 1993) is a powerful method for market basket analysis, which aims at finding regularities in the shopping behavior of customers. An association rule is a rule like "Z <= X, Y." The items X and Y are called antecedents in the rule and Z is the consequent. This rule expresses an association between items X, Y, and Z. It states that if a customer is picked randomly and the customer selected items X and Y, it is likely that the customer also selected item Z. The number of antecedents can range from one to the total number of items in a database.

Initially, two parameters are reported for association rules, namely the support, which estimates the probability P({X,Y,Z}), and the confidence, which estimates the probability P(Z|{X, Y}). An item set {X, Y, Z} is large (or frequent) if its support is greater than or equal to the user specified minimum support. An association rule Z <= X, Y is strong if it has a large support and a high confidence. The so called "support-confidence framework" for association rules does not contain information for negative implications. In other words, we cannot figure out how likely it is that a customer purchasing X and Y does not purchase Z. The third parameter, the lift (Silverstein et al. 1998), measuring the strength of a rule "Z<= X, Y," is introduced as the ratio between the actual probability of the item set containing both antecedent and consequent divided by the product of the individual probabilities of the antecedent set and the consequent. That is,

$$\text{lift} = P(\{X,Y,Z\})/[P(\{X,Y\})*P(Z)].$$

The lift measures the dependency among the antecedents and the consequent of a rule. Lift values above one indicate positive dependence, while those below one indicate negative dependence.

The version of the association rule algorithm we used in this study is implemented by Borgelt [2007]. The support value in this implementation is defined as P({X,Y}) instead of P({X,Y,Z}). Since no previous experience exists on this specific data mining task for RI hurricanes, no specific control parameters (predefined support, confidence, and lift) are chosen. Instead, results are compared to the results of KD03, and the rules given significant results are discussed.

Generally speaking, 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., Steinbach et al. 2003). The other approach is to convert the scientific data into transaction data and then use classical data mining tools such as association rules to reveal hidden relationships between different physical parameters. In this study, since we use a traditional data mining tool, we need to discretize 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. In other words, the antecedent in this case are the 22 input features (11 variables with the "high" and "low" discretized values) and the consequent is "FD24=RI." 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

KD03 found that higher RI probabilities can be obtained when a combination of conditions is satisfied and identified that the highest RI probabilities occurred when five conditions (high PD12, low SHRD, high SST, high POT, and high RHLO) are satisfied together. They found that about 2% of the total samples satisfy this condition combination and among them 41% underwent RI.

The above cited result can be easily translated into an association rule as "RI <= SHRD=L, PD12=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 differences in the exact numbers are likely caused by the differences in data selections because we did not include non-developing depressions and did not consider the land factor. The differences resulted in a 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⇐ SHRD=L, PD12=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⇐ SHRD=L, PD12=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, SHRD=L, PD12=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 accuracy in determinating RI cases. Meanwhile, without proper guidance, it is difficult to select RI constraints in the RI probability estimate to achieve the improved results. The association rule mining with the pruning process provides a systematic implementation to reach this goal.

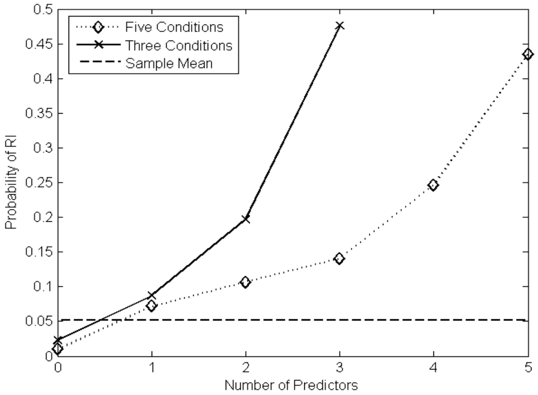


Figure 1. The composite RI probabilities based on the condition set in KD03 {PD12=H, SST=H, POT=H, RHLO=H, SHRD=L} and the three-mined-condition set {PD12=H, RHLO=H, SHRD=L}. The sample mean RI probability is also shown for reference.

To compare the results with those in KD03, the composite RI probability defined in KD03 is computed. 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 when all five constraints identified in KD03 satisfied. The RI probability when all five KD03 conditions are 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 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 (Hand et al. 2001). The data mining results also shed light for potential improvement of TC intensity forecasting.

One more step further, we can continue the search for the optimal condition for a given number of satisfied conditions. For example, we can limit the number to three and search the optimal conditions, which gives the highest RI probability in all combinations with three conditions. Surprisingly, the result is different from what we found inside the five condition range selected by KD03. The mined optimal combination with three conditions are “U200=L, PD12=H, USTM=H, *(supp=0.8%, conf=48.1%, lift=941.9%).*”

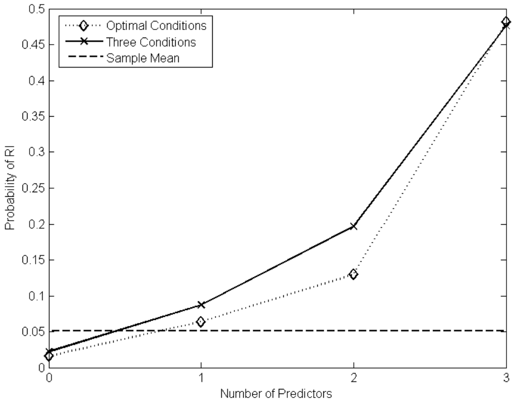


Figure 2. The composite RI probabilities based on the two searched sets with three conditions, the optimal set {U200=L, PD12=H, USTM=H} and the set compared with the RI probability of KD03 {PD12=H, RHLO=H, SHRD=L}. The sample mean RI probability is also shown for reference.

Figure 2 shows the comparison of the two cases. We can find that the difference is not significant. Although when the condition number is three, the “optimal” case that gives the highest RI probability, the difference is not large (48.1% vs. 47.6%). Actually, if we check the details in Figure 2 for the cases with one and two conditions in each group, we find that the original case gives better results for both conditions. It “lost” to the optimal condition only when we consider all three conditions together. Therefore, for this specific case with three conditions, we would say that the subset from the five conditions given in KD03 is almost the ‘optimal” set.

This is a relatively surprising result because among the three members, two of them are different from what we obtained for comparing with the results of KD03. It is suspected that the TC cases associated with the two cases may come from different TC samples. Detailed study is needed to understand the differences.

5. CONCLUSION

Researchers have developed various methods to predict the intensity of TCs (DeMaria et al. 2007). Many of those methods suffer degradation of performance for rapidly intensifying tropical cyclones (DeMaria and Kaplan 1994). In this study, we applied the association rule to look for combinations of persistent and synoptic conditions which provide improved RI probability estimates.

Compared to statistical analysis, the technique of association rules can explore associations among multiple conditions with little 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 screen those conditions more comprehensively with a 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 volumes is not a trivial task. A better cyber-infrastructure with intelligent archiving features such as server-side data manipulation and data aggregation may significantly reduce the time-demanding preprocessing load. One such potential cyber-infrastructure for this is the GrADS Data Sever (Wielgosz et al. 2001).

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

Agrawal, R., T. Imielinski, and A. Swami (1993), Mining association rules between sets of items in large databases, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington D.C., May 1993, 207-216.

Borgelt, C. (2007), Apriori - Association Rule Induction / Frequent Item Set Mining, http://www.borgelt.net/apriori.html, last access on May 24, 2007.

DeMaria, M. (1996), The effect of vertical shear on tropical cyclone intensity change, *J. Atmos. Sci.*, *53*, 2076-2087.

DeMaria, M. and J. Kaplan (1999), An Updated Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic and Eastern North Pacific Basins Mark, *Weather and Forecasting,* 14, 326–337. DOI: 10.1175/1520-0434(1999)014<0326:AUSHIP>2.0.CO;2

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

DeMaria, M., J.A. Knaff, and C.R. Sampson, 2007: Evaluation of Long-Term Trends in Operational Tropical Cyclone Intensity Forecasts. *Meteor. and Atmos. Phys.*, **59**, 19-28.

DeMaria, M., M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan (2005), Further Improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS), *Weather* *and Forecasting, 20*, 531–543. DOI: 10.1175/WAF862.1.

Gray, W. M. (1968), Global view of the origin of tropical disturbances and storms, *Mon. Wea. Rev.*, *96*, 669–700.

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

Holliday, C. R., and A. H. Thompson (1979), Climatological characteristics of rapidly intensifying typhoons. *Mon. Wea. Rev*., 107, 1022–1034.

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

Kaplan, J. and M. DeMaria (2003), Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin, *Weather and Forecasting*, *18*, 1093-1108.

Lee, W.-C., and M. M. Bell (2007), Rapid intensification, eyewall contraction, and breakdown of Hurricane Charley (2004) near landfall, *Geophys. Res. Lett*., *34*, L02802, doi:10.1029/2006GL027889.

Merrill, R. T. (1988), Environmental influences on hurricane intensification, *J. Atmos. Sci.*, *45*, 1678–1687.

Silverstein, C., S. Brin, and R. Motwani (1998), Beyond market baskets: generalizing association rules to dependence rules. Data Mining and Knowledge Discovery, 2*, 39-68.*

Steinbach, M., Tan, P. N., Kumar, V., Potter, C. and Klooster, S, 2003. Discovery of Climate Indices using Clustering. The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Wang, Y. (2002), An Explicit Simulation of Tropical Cyclones with a Triply Nested Movable Mesh Primitive Equation Model: TCM3. Part II: Model Refinements and Sensitivity to Cloud Microphysics Parameterization *Mon. Wea. Rev.*, *130*, 3022–3036. DOI: 10.1175/1520-0493(2002)130<3022:AESOTC>2.0.CO;2

Wielgosz, J., B. E. Doty , J. Gallagher, and D. Holloway, “GrADS and DODS,” 17th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Jan 2001.

Willoughby, H. E., J. A. Clos, and M. G. Shoreibah (1982), Concentric eyewalls, secondary wind maxima, and the evolution of the hurricane vortex, *J. Atmos. Sci.*, *39*, 395–411.