Open problem: Dynamic Relational Models for Improved Hazardous Weather Prediction

Amy McGovern[1] AMCGOVERN@OU.EDU
[1]School of Computer Science, University of Oklahoma, Norman, OK 73019 USA

Adrianna Kruger[1] ADRIANNAKRUGER@OU.EDU
Derek Rosendahl[2] DROSE@OU.EDU
[2]School of Meteorology, University of Oklahoma, Norman, OK 73019 USA

Kelvin Droegemeier[2] KKD@OU.EDU

1. Introduction

We are developing dynamic relational knowledge discovery methods for use on mesoscale weather data. Severe weather phenomena such as tornados, thunderstorms, hail, and floods, annually cause significant loss of life, property destruction, and disruption of the transportation systems. The annual economic impact of these mesoscale storms is estimated to be greater than $13B (Pielke and Carbone, 2002). Any mitigation of the effects of these storms would be beneficial. However, current techniques for predicting severe weather are tied to specific characteristics of the radar systems. Each new sensing system requires the development of new radar detection algorithms for detecting hazardous events. Our research focuses on developing new dynamic relational models that will enable meteorologists to improve their understanding of the formation of tornados and other severe weather events.

Current weather radar detection and prediction systems primarily rely on numerical models. We propose to enhance our understanding of the formation of severe weather events, specifically focusing on tornados, through knowledge discovery. The process of knowledge discovery is about making sense of data. Generally, the data are too complex for humans to quickly understand and identify the important patterns. Instead, knowledge discovery techniques can be used to highlight salient patterns.

Instead of viewing the data through physical equations, we will view the data at a higher level. Weather forecasters identify high-level features in the radar readings and use the relationships among these features to determine whether or not a severe storm is building. For example, seeing a hook echo inside a supercell thunderstorm is a sign that a tornado is likely to occur in the near future. Many machine learning methods focus on a propositional representation. However, allowing algorithms to reason about high-level entities and relationships facilitates the discovery of more complicated patterns. A relational representation provides a richer language and it aligns closely with the representation already used by the human weather forecasters.

2. Meteorological data

Figure 1 (left panels) shows a sample of the four dimensional gridded data that will be made available via ensemble Kalman filter (EnKF) assimilation of real observations (Tong and Xue, 2005). The leftmost panels show the radar retrieved wind velocities and reflectivity observed during the May 29, 2004 tornado in Oklahoma City. The remaining panels show assimilated radar reflectivity, temperature deviation, pressure deviation, and vertical vorticity.

The data available for a knowledge discovery system are comprised of three-dimensional cubes, or voxels, each of which has a set of meteorological variables associated with it. The fourth dimension of the gridded data is time. Fundamental variables include the x, y, z coordinates, wind readings in each direction, precipitation content of a cube, temperature, and pressure. Examples of the derived variables include divergence (spreading out of winds), temperature gradient, and the pressure gradient force.

Each voxel has a set of readings for each of the meteorological variables. These readings change over time,

Presented at the ICML Workshop on Open Problems in Statistical Relational Learning, Pittsburgh, PA, 2006.

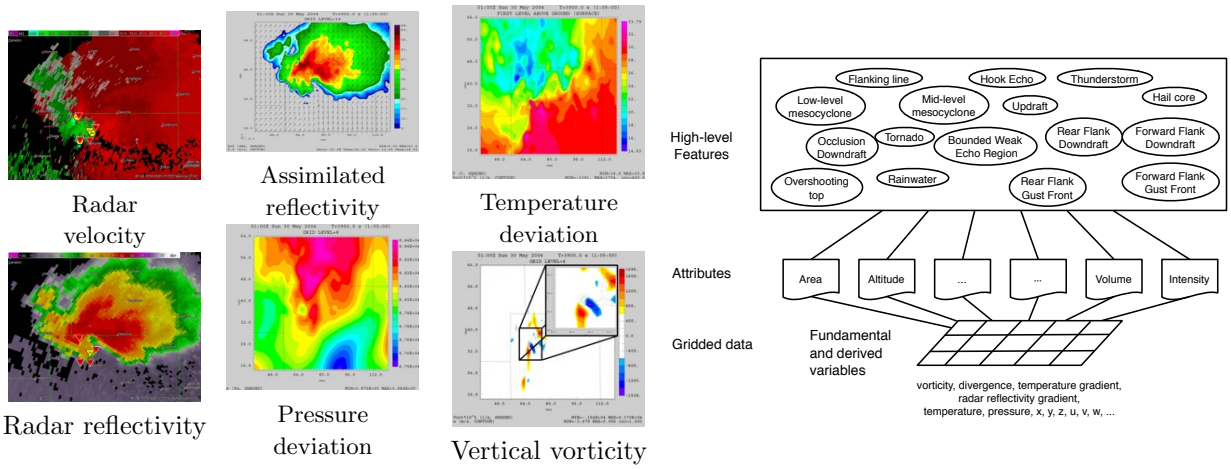

Figure 1. An example of real data (left panels) being used to create gridded data (center and right panels). This data is from the May 29, 2004 tornado in Oklahoma City and is courtesy of Fritchie and Droegemeier at the University of Oklahoma.

giving us a large and dynamic data set. For example, a reasonable storm simulation may be 100km by 100km in area by 18km in height. Although the grid spacing is generally not equal throughout the region, it can be as small as 100m. This spacing can be reduced at the expense of computation time. Recorded every 30 to 60 seconds, data for a single storm quickly becomes overwhelming.

This four-dimensional gridded data will be available from both assimilated weather observations and simulation. Droegemeier et al. (2005) is developing a web-based computing infrastructure that will support a more general approach to severe weather detection and prediction. The technology to assimilate the data into gridded form in real time is under active development. We can also use simulated storm data produced from the Advanced Regional Prediction System, which is one of the top weather forecasting systems for mesoscale data (Xue et al., 2003).

Figure 1 (right panel) gives an overview of how we will use the gridded data. For the purposes of knowledge discovery, the gridded data will be observed through its attributes. Observing the attributes enables us to identify the high-level features. The set of features in this figure represent our current set for predicting tornados.

3. Key challenges and benefits

Mesoscale meteorological data provide a number of challenges for knowledge discovery techniques. Given the anticipated difficulty of working with this data, the key reason to study it is the potential for significant and tangible benefits.

The primary challenge comes from the temporal nature of the data. There are data mining techniques for dynamic data but these approaches use a propositional representation (for example, see Zaki, 2001; Zaki et al., 2005; Oates and Cohen, 1996). Although we are currently using a propositional approach, we believe that a relational approach will enable us to better understand the formation of tornados. In addition to dynamic, the data are continuous and multi-dimensional. Even with a propositional representation, identifying patterns in continuous data is difficult. We are currently using Lin et al.'s (2003) approach to creating discrete data from continuous data. The multi-dimensional aspect to the problem only makes it more challenging. There is recent work addressing this issue (for example, see Tanaka and Uehara, 2003) but how to best mine multi-dimensional time series is still an open problem.

Although weather forecasters make use of high-level features such as those presented in Figure 1, it is difficult to automatically extract this data. For example, defining the exact boundaries of a storm is not straightforward. Does a storm end where the rain ends completely or where it is only drizzling? Are the wind speeds important? Perhaps the most difficult aspect of answering these questions is defining an answer that a majority of meteorologists will agree upon. Related to this aspect, if we had the ability to identify all of the high-level features, a relational approach would then need to identify the relationships among these features. Since the features are four dimensional, an exhaustive search for possible relationships is not feasible.

We are not aware of any statistical relational models that can handle the volume of dynamic relational data presented here. Applying statistical relational knowledge discovery tools to this data requires the development of both a new dynamic relational data representation and new dynamic relational models. Although relational representations have been proven successful in many real-world data mining examples with dynamic data (e.g., Kubica et al. (2003); McGovern et al. (2004); Neville et al. (2005)), these representations ignore the temporal aspect of the data. A first order logical representation such as Dzeroski (1995) or Blockeel and DeRaedt (1998) cannot express stochastic or dynamic data. This cripples it for use on the weather data. Although Richardson and Domingos's (2005) approach can handle stochastic data, it cannot deal with temporal data. Sanghai et al. (2003) is the only example of a principled approach to dynamic relational modeling that we are aware of. Their work combines a dynamic bayes net (DBN) approach with the PRM approach, where the PRM is used to represent data at a single moment in time and the DBN specifies how the relations may change over time.

The biggest potential for benefit will come from an increased understanding of how tornados form, thus improving their predictability. Our specific long-term goals include reducing the number of false positives (currently about 75%) and increasing the lead time for warnings (currently about 12 minutes). Simmons and Sutter (2005) recently demonstrated that Doppler radars save an average of 80 lives per year when the warning lead time is increased by only a few minutes.

A key benefit of developing new prediction techniques based on assimilated data is that the creation of new radar or other sensing systems does not necessitate the development of new detection and prediction techniques. Instead of changing the form of the data that is available, new sensing systems improve the quality of the gridded analysis, leading to improved prediction. From a computer science perspective, any new models or data representations developed for use on this data will have to work on actual assimilated weather data.

References

Blockeel, H. and DeRaedt, L. (1998). Top-down induction of first order logical decision trees. *Artificial Intelligence*, 101(1-2):285–297.

Droegemeier, K. K., Baltzer, T., Brewster, K., Clark, R., Domenico, B., Gannon, D., Graves, S., Joseph, E., Morris, V., Murray, D., Plale, B., Ramachandran, R., Ramamurthy, M., Ramakrishnan, L., Reed, D., Rushing, J., Weber, D., Wilhelmson, R., Wilson, A., Xue, M., and Yalda, S. (2005). Service oriented environments in research and education for dynamically interacting with mesoscale weather. *Computing in Science and Engineering*, 7:12–29.

Dzeroski, S. (1995). Inductive logic programming and knowledge discovery in databases. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, *Advances in Knowledge Discovery and Data Mining*, pages 118–152.

Kubica, J., Moore, A., and Schneider, J. (2003). Tractable group detection on large link data sets. In Wu, X., Tuzhilin, A., and Shavlik, J., editors, *The Third IEEE International Conference on Data Mining*, pages 573–576. IEEE Computer Society.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*.

McGovern, A., Friedland, L., Hay, M., Gallagher, B., Fast, A., Neville, J., and Jensen, D. (2004). Exploiting relational structure to understand publication patterns in high-energy physics. *SIGKDD Explorations*, 5(2):165–172. Winning entry to the open task for KDD Cup 2003.

Neville, J., Şimşek, Ö., Jensen, D., Komoroske, J., Palmer, K., and Goldberg, H. (2005). Using relational knowledge discovery to prevent securities fraud. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page To appear.

Oates, T. and Cohen, P. R. (1996). Searching for structure in multiple streams of data. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 346–354. Morgan Kauffman.

Pielke, R. and Carbone, R. (2002). Weather impacts, forecasts, and policy. *Bulletin of the American Meteorological Society*, 83:393–403.

Richardson, M. and Domingos, P. (2005). Markov logic networks. *Machine Learning*, To appear.

Sanghai, S., Domingos, P., and Weld, D. (2003). Dynamic probabilistic relational models. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. Morgan Kauffman.

Simmons, K. M. and Sutter, D. (2005). Wsr-88d radar, tornado warnings, and tornado casualties. *Weather and Forecasting*, 20(3):301–310.

Tanaka, Y. and Uehara, K. (2003). Discover motifs in multi-dimensional time-series using the principal component analysis and the mdl principle. In *Proceedings of the Third International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2003)*, pages 252–265.

Tong, M. and Xue, M. (2005). Ensemble kalman filter assimilation of doppler radar data with a compressible nonhydrostatic model: Oss experiments. *Mon. Wea. Rev.*, 133:1789–1807.

Xue, M., Wang, D., Gao, J., Brewster, K., and Droegemeier, K. K. (2003). The advanced regional prediction system (arps), storm-scale numerical weather prediction and data assimilation. *Meteorology and Atomospheric Physics*, 82:139–170.

Zaki, M. J. (2001). Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1/2):31–60. Special issue on unsupervised learning.

Zaki, M. J., Parimi, N., De, N., Gao, F., Phoophakdee, B., Urban, J., Chaoji, V., Hasan, M. A., and Salem, S. (2005). Towards generic pattern mining. In *International Conference on Formal Concept Anaysis*.