Real-Time Mining of Integrated Weather Data

Workshop in Dynamic Data Driven Applications
ICCS 2004
Krakow, Poland

Presenter: Theodore Trafalis
Collaborators: A. White, M. Richman, P. Skubic, V. DeBrunner, V. Lakshmanan, S. Lakshmivarahan
This project has been lasting for almost two years. The 3rd year will start from August 2004, funded by NSF.
Severe Weather Forecasting

- Why are we doing this?
  - Improve warning times
  - Accuracy of severe weather forecasts

- Techniques ↔ Domain Specific Activities/Knowledge
  - Pattern Recognition
  - Machine Learning
  - Data Mining
  - Multi Sensor Data Fusion
  - Severe Weather Detection
  - Weather Radar Post Processing
Goals of the project

- Techniques to mine massive amounts of weather data in real-time
- Multi-sensor data fusion of radar products to detect severe weather events
- Detection of severe weather signatures
- Statistical learning techniques applied to weather data
Techniques to mine massive amounts of weather data in real-time

- Data reduction

- Summary of “Tornado detection algorithm using Empirical Orthogonal Functions” P. Kakani, S. Lakshmivarahan, M. Richman. Submitted to ANNIE 2004
MDA Attributes

- 1. base (m) [0-12000]
- 2. depth (m) [0-13000]
- 3. strength rank [0-25]
- 4. low-level diameter (m) [0-15000]
- 5. maximum diameter (m) [0-15000]
- 6. height of maximum diameter (m) [0-12000]
- 7. low-level rotational velocity (m/s) [0-65]
- 8. maximum rotational velocity (m/s) [0-65]
- 9. height of maximum rotational velocity (m) [0-12000]
- 10. low-level shear (m/s/km) [0-175]
- 11. maximum shear (m/s/km) [0-175]
- 12. height of maximum shear (m) [0-12000]
- 13. low-level gate-to-gate velocity difference (m/s) [0-130]
- 14. maximum gate-to-gate velocity difference (m/s) [0-130]
- 15. height of maximum gate-to-gate velocity difference (m) [0-12000]
- 16. core base (m) [0-12000]
- 17. core depth (m) [0-9000]
- 18. age (min) [0-200]
- 19. strength index (MSI) wghtd by avg density of integrated layer [0-13000]
- 20. strength index (MSIr) "rank" [0-25]
- 21. relative depth (%) [0-100]
- 22. low-level convergence (m/s) [0-70]
- 23. mid-level convergence (m/s) [0-70]
## Confusion matrix and forecast indices

### Confusion matrix

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Observed</th>
<th>Forecast Yes</th>
<th>Forecast No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Hits (a)</td>
<td>False alarm (b)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Misses (c)</td>
<td>Correct negative (d)</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Observed yes</td>
<td>Observed No</td>
<td></td>
</tr>
</tbody>
</table>
The performance of the techniques used is assessed by using the following forecast evaluation indices:

- Probability of Detection (POD) = \[
\frac{a}{a + c}
\]

- False Alarm Rate (FAR) = \[
\frac{b}{a + b}
\]

- Critical Success Index (CSI) = \[
\frac{a}{a + b + c}
\]

- Bias = \[
\frac{a + b}{a + c}
\]

- Probability of false Detection (POFD) = \[
\frac{b}{b + d}
\]

- Heidke’s Skill = \[
2(ad-bc)/((a+b)(b+d)+(a+c)(c+d))
\]
EOF Analysis - Maximum eigen value component
Discrimination Rules

**Step 1:** Let $p_1$, $p_2$, $p_3$ be projections of an input vector $y$ with 23 components on the first three eigenvectors for tornados and $q_1$, $q_2$, $q_3$ be those for non-tornados.

**Step 2:** IF \{ $(p_1 \geq 10,000) \& (q_2 < 2500) \& (p_3 < -1500) \text{OR} (p_3 > 1000)$\} 
THEN

y is a tornado.

ELSE

y is a non-tornado.
Conclusions

- EOF algorithm was useful in the discrimination between tornadoes and non-tornadoes
- 3 components extract 95% of variance contained in the original variables
- EOF reduces the amount of overlap
- Decision rules were made to be simple, with only three components and achieve over 95% for POD, the highest among all the known methods
Vignettes of dynamic learning algorithms

- Decision tree paper
- SVM with including month
- ANNIE 2004
Decision Tree Diagram of rule generation and decisions. Ovals represent nodes, squares represent leaves.

- X1 < 90?
  - No
  - Yes

- X2 < 617?
  - No
  - Yes

- X2 > 12219?
  - No
  - Yes

- X3 < 0?
  - No
  - Yes

- X3 > 13?
  - No
  - Yes

- X23 > 28?
  - No
  - Yes

- SVM

- NT

- T

- unclassified
Boxplots of misclassification error due to (a) non-tornado rules set
Boxplots of misclassification error due to (b) SVM
Boxplots of misclassification error due to (c) tornado rules set and
Boxplots of misclassification error due to (d) total hybrid system.
Misclassification error of the hybrid system components and total system (top) and Misclassification error for SVM (bottom)

<table>
<thead>
<tr>
<th>Multiplier</th>
<th>0.90</th>
<th>0.95</th>
<th>1.00</th>
<th>1.05</th>
<th>1.10</th>
<th>1.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-tornado rules</td>
<td>0.0191</td>
<td>1.00</td>
<td>0.0174</td>
<td>0.0454</td>
<td>0.086</td>
<td>0.1373</td>
</tr>
<tr>
<td>Tornado rules</td>
<td>0.0316</td>
<td>0.0237</td>
<td>0.0550</td>
<td>0.0968</td>
<td>0.1032</td>
<td>0.1550</td>
</tr>
<tr>
<td>SVM</td>
<td>0.2024</td>
<td>0.0174</td>
<td>0.2110</td>
<td>0.1997</td>
<td>0.2093</td>
<td>0.2154</td>
</tr>
<tr>
<td>Total system</td>
<td>0.1648</td>
<td>0.0474</td>
<td>0.1648</td>
<td>0.1449</td>
<td>0.1485</td>
<td>0.1725</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SVM</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
<th>Sample 5</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1664</td>
<td>0.1778</td>
<td>0.1713</td>
<td>0.1811</td>
<td>0.1566</td>
<td>0.1706</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

- The best misclassification error for the hybrid model (0.1449) is 12.7% lower than the one for the model based solely on SVM (0.1706).
Box plots of the seasonal variability of the mean of 23 attributes.
Misclassification rates for SVM(top), MPM(middle), NN(bottom).

<table>
<thead>
<tr>
<th>Ratio of tornado-no tornado</th>
<th>2%</th>
<th>4%</th>
<th>6%</th>
<th>8%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without month</td>
<td>0.017</td>
<td>0.025</td>
<td>0.033</td>
<td>0.037</td>
<td>0.043</td>
</tr>
<tr>
<td>With month</td>
<td>0.013</td>
<td>0.021</td>
<td>0.027</td>
<td>0.032</td>
<td>0.037</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ratio of tornado-no tornado</th>
<th>2%</th>
<th>4%</th>
<th>6%</th>
<th>8%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without month</td>
<td>0.013</td>
<td>0.025</td>
<td>0.033</td>
<td>0.038</td>
<td>0.044</td>
</tr>
<tr>
<td>With month</td>
<td>0.012</td>
<td>0.020</td>
<td>0.026</td>
<td>0.031</td>
<td>0.036</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ratio of tornado-no tornado</th>
<th>2%</th>
<th>4%</th>
<th>6%</th>
<th>8%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without month</td>
<td>0.079</td>
<td>0.078</td>
<td>0.088</td>
<td>0.089</td>
<td>0.092</td>
</tr>
<tr>
<td>With month</td>
<td>0.099</td>
<td>0.099</td>
<td>0.110</td>
<td>0.114</td>
<td>0.114</td>
</tr>
</tbody>
</table>
POD: measures the percentage of tornadoes detected
FAR: measures false alarm
Skill: how much a forecast is superior to some known reference forecast (e.g., random chance).
Conclusions

- Adding month number (1,2,..,12) as an attribute has improved the results for SVM and MPM but not for NN.
- Overall, the SVM technique has a slight edge over the MPM method for most forecast evaluations. However, the key finding is that MPM and SVM are more accurate, have a lower false alarm rate, less bias and more skill than the NN technique.
Objectives

- Support Vector Machines (SVM) and Minimax Probability Machines (MPM), fueled by dynamically driven data, are applied to the tornado detection problem.
- The inputs to the SVM and MPM are from the National Severe Storms Laboratory (NSSL) Mesocyclone Detection Algorithm (MDA).
- As each input varies, as a function of time, we investigate the behavior of the time signals before and the time of the tornado touch down to seek the dynamic pattern that gives rise to such storms.
Forecast evaluation indices using SVM for different time lags

<table>
<thead>
<tr>
<th>Time lag</th>
<th>[t-2]</th>
<th>[t-2,t-1]</th>
<th>[t-2,t-1,t]</th>
<th>[t-1]</th>
<th>[t-1,t]</th>
<th>[t]</th>
</tr>
</thead>
<tbody>
<tr>
<td>POD</td>
<td>0.6417</td>
<td>0.6204</td>
<td>0.6515</td>
<td>0.5908</td>
<td>0.6451</td>
<td>0.6341</td>
</tr>
<tr>
<td>FAR</td>
<td>0.3637</td>
<td>0.357</td>
<td>0.3036</td>
<td>0.4078</td>
<td>0.3381</td>
<td>0.3688</td>
</tr>
<tr>
<td>POFD</td>
<td>0.3623</td>
<td>0.3403</td>
<td>0.2807</td>
<td>0.4019</td>
<td>0.3256</td>
<td>0.366</td>
</tr>
<tr>
<td>Bias</td>
<td>1.0085</td>
<td>0.9648</td>
<td>0.9356</td>
<td>0.9976</td>
<td>0.9746</td>
<td>1.0046</td>
</tr>
<tr>
<td>CSI</td>
<td>0.4695</td>
<td>0.4615</td>
<td>0.5074</td>
<td>0.42</td>
<td>0.4852</td>
<td>0.4627</td>
</tr>
<tr>
<td>Heidke’s Skill</td>
<td>0.2793</td>
<td>0.2802</td>
<td>0.371</td>
<td>0.1889</td>
<td>0.3195</td>
<td>0.2681</td>
</tr>
</tbody>
</table>
Forecast evaluation indices using SVM by adding the output from previous time to the next consecutive time

<table>
<thead>
<tr>
<th>Time Lag</th>
<th>[t-2,t-1,t]</th>
<th>[t-1,t]</th>
</tr>
</thead>
<tbody>
<tr>
<td>POD</td>
<td>0.6381</td>
<td>0.5879</td>
</tr>
<tr>
<td>FAR</td>
<td>0.361</td>
<td>0.4054</td>
</tr>
<tr>
<td>POFD</td>
<td>0.3562</td>
<td>0.3961</td>
</tr>
<tr>
<td>Bias</td>
<td>0.9986</td>
<td>0.9888</td>
</tr>
<tr>
<td>CSI</td>
<td>0.4691</td>
<td>0.4197</td>
</tr>
<tr>
<td>Heidke’s Skill</td>
<td>0.282</td>
<td>0.1918</td>
</tr>
</tbody>
</table>
Conclusions

- The dynamic inclusion of data has a beneficial result on tornado prediction.
- As the lead-time to prediction is increased, the accuracy of the forecast decreases. This is an expected result.
Vortex Identification

- Lakshmanan, Smith
- Kshirsagar, Skubic
- White, Decker
Vortex Detection and Diagnosis (VDDA)

- Linear-Least Squares Derivatives (LLSD) of velocity
  - Rotation and Divergence
- Multi-radar mosaic
Vortex Detection and Diagnosis (VDDA)

- Linear-Least-Squares Derivatives (LLSD) of velocity
  - Rotation and Divergence
- Multi-radar mosaic
Detection of Severe Weather Signatures

- Image Processing methods to detect circulation patterns for the prediction of Tornado Vortex Signatures and Mesocyclones

- Summary of “Detection of Mesocyclones & Tornado Vortex Signatures (TVS) using Pattern Recognition Techniques”
  
  P. Kshirsagar, P. Skubic.

  To be submitted to AIPR 2004
Observed Patterns
Parameters Processed - Azimuthal Shear & Reflectivity

Reflectivity Signature
HOOK

Azimuthal Shear Signature
TRIPLET
Approach

- Blob Coloring of Reflectivity Data to isolate high values
- Watershed Segmentation ofAzimuthal Shear data to detect patterns
Performance Analysis

- Testing & Training of 24 Tornado & 24 Non-Tornado Cases
- Critical Success Index $= 0.32 \pm 0.07$
- Probability of Detection $= 0.96 \pm 0.27$
- False-Alarm Rate $= 0.68 \pm 0.15$
- Lead time of over 20 min. achieved for violent (F-3 to F-5) tornadoes
Shear patterns in Level II Doppler radar data near severe tornadic thunderstorms

- Current circulation detection algorithms ignore anticyclonic shear information (Stumpf et al. 1998)
- Search near storm environment for anticyclonic shear patterns adjacent to cyclonic circulations
- Discovered clear discriminator
  - Tripole pattern persistent in severe (F-3,F-4,F-5) tornadic thunderstorms
  - Tripole pattern absent in weak thunderstorms

(Donaldson, 1970)
Current Findings

- Tripole discriminator pattern
  - Tripole structure present in all test cases \((n=12)\) of F-5, F-4 tornado events and most F-3 tornado events
  - Tripole structure absent in all test cases \((n=50)\) of non-tornadic and F-0, F-1, F-2 tornado events

- Tripole discriminator pattern attributes are functions of...
  - Mesocyclone range from radar
  - Mesocyclone orientation w.r.t. radar
  - Mesocyclone rotational strength
Example of tripole discriminator pattern in mesocyclone that produces F-4 tornado
Gust front detection

- DeBrunner, Alkhouli
The objective is to develop image processing methods that identify gust fronts in reflectivity and wind shear radar images.

The above objective falls under edge-detection or template matching techniques.

Template matching offers more flexibility and has recently been under great investigation. Template matching can be performed by the following techniques:

- Standard cross correlation
- Mathematical Morphologies
- Functional template correlation. Delanoy et. al.

Recently our group implemented Delanoy method to identify boundaries in reflectivity and shear radar images and fuse them into one more accurate boundary image.
We have modified Delanoy’s method by using Entropy as a measure of the similarity between the image and the template.

Entropy of any histogram $H(i)$ is defined as follows:

$$E = - \sum_i H(i) \log H(i)$$

The features in the template can be encoded into a histogram and the entropy of this histogram can be used as a template matching technique.

Entropy is a measure of the information in images. Entropy measure is invariant under changing the orientation of the features in the template and the image.
Due to these properties, template matching with entropy is performing better.

This fact is evident from the figures above. The boundaries in the second image are clear and more connected.
Multiple Radar

- Lakshmanan, Smith
Multiple Radar SSAP

- Runs traditional Severe Storms Analysis Package (SSAP) algorithms using multiple radar input and Near Storm Environment (NSE) data.
- Can use adjacent radars to fill cones-of-silence.
- Outputs information rapid intervals (60 second updates); can be as fast as individual elevation scan updates using "virtual volume scans".
  - "Rapid update" also works in single-radar mode if coverage or outages dictate.
Multiple Radar SSAP
MR-SSAP Rapid Update

- 20 May 2000
- KSRX (Fort Smith)
Combination Strategies

- Several strategies exist to combine data from multiple radars:
  - Distance-weighting
  - Time-weighting
  - Advection of older data
  - Maximum
  - Absolute-value maximum
  - Combinations of these (if not contradictory)
- Users choose combination strategy based on the type of data being merged.
- Advection based on motion estimates.
Time-Distance weighting
Time-Distance + advection
Summary

- MPM and SVM applications can improve the performance of skill of tornado forecast compared to MDA and ANN. Adding more time steps using a DDDAS approach to predict tornadoes improves the skill.

- Multi Sensor Data Fusion using multiple radars has allowed for improved detection of tornadic signatures and much faster scanning of the atmosphere.

- In the past, tornado warning time were on the order of 8 minutes. As a result of all of our analyses, signatures of tornadoes can be detected on average with 24 minutes lead time using pattern recognition and data mining. For strong tornadoes, lead time is up to 30 minutes. This should lead to many more lives saved.