AUTOMATED CLASSIFICATION OF CONVECTIVE AREAS IN REFLECTIVITY USING DECISION TREES

David John Gagne II
Center for Collaborative Adaptive Sensing of the Atmosphere Research Experience for Undergraduates
University of Oklahoma
Norman, OK

Amy McGovern
School of Computer Science
University of Oklahoma
Norman, OK

Jerry Brotzge
Center for the Analysis and Prediction of Storms
University of Oklahoma
Norman, OK

ABSTRACT

This paper presents an automated approach to classifying storms based on their structure using decision trees. When dealing with large datasets, manually classifying storms quickly becomes a repetitive and time-consuming task. An automated system can more quickly and efficiently sort through large quantities of data and return value-added output in a form that can be more easily manipulated and understood. Our method of storm classification combines two machine learning techniques, k-means clustering and decision trees. K-means segments the reflectivity data into clusters and decision trees classify each cluster. We chose decision trees for their simplicity and ability to screen out unimportant attributes.

We used a k-means clustering algorithm derived from Lakshamanan (2001) to divide the reflectivity into different regions. Each cluster was sorted as convective or stratiform based on reflectivity. Each convective cluster was hand labeled at both a general and a specific level. The two general classifications were storm cells and linear systems. The specific classifications for cells were isolated severe, isolated non-severe, and circular Mesoscale Convective System (MCS). The specific classifications for linear systems were trailing stratiform, leading stratiform, and no or parallel stratiform. We used the Waikato Environment for Knowledge Analysis (WEKA), a machine learning suite, to develop the decision trees (Witten and Frank, 2005).

We constructed multiple decision trees with both morphological and reflectivity attributes for both the general and specific classifications. The training and test data sets came from Advanced Regional Prediction System (ARPS) simulated reflectivity data (Xue et al., 2001, 2002, 2003), and we created an additional data set from a collection of composite reflectivity mosaics from the CASA IP1 network (Brotzge et al., 2006). Overall, the best accuracy for the general type trees stayed in the 90% range for all three test sets indicating a very reliable classification tree. By verifying the trees learned on simulated data with observations from the CASA network, we demonstrated that the knowledge gained from simulation can be applied to real situations. For the specific type, the accuracy ranged from 55% to 80% across the test sets, implying additional work is needed for improvement.

1. INTRODUCTION

The taxonomy of storm classification presents many challenges even for human experts. The nature of a classification system changes depending on how the storm is observed (Doswell et al, 1996). The definition of a mesoscale convective complex, for instance, requires infrared satellite readings (Maddox, 1982). The limits of the observational tool also determine what storm types can be identified. For example, researchers focusing their classification systems on rainfall measurements (Baldwin et al, 2004), and radar returns (Steiner et al., 1995; Biggerstaff et al., 2000; Rigo and Llasat, 2004; Anagnostou, 2004), ignore a type like MCC and only use types visible from their specific data source. This project shares that limitation with previous work in its own focus on one variable – reflectivity – from one instrument, weather radar.

An automated system provides numerous advantages over manual classification. When dealing with large datasets on the order of thousands or more storms, manual extraction from a data set is impossible in a reasonable amount of time. With an automated system, however, an algorithm can quickly and efficiently iterate through the data processing information as needed. Without the burden of hand labeling, researchers can spend more time on data analysis.

Steiner, Houze, and Yuter (1995, SHY95) employed a technique that separated convective and stratiform areas using a combination of

intensity and difference from background reflectivity (1995), a variation on the background-exceedence technique (Biggerstaff et al., 2000). Anagnostou devised a different approach to the convective/stratiform divide by employing neural networks as a means to form multiple parameters into a separation function (2004). Part of the proposed methodology requires a separation of convective and stratiform areas, but instead of searching for the convective areas and labeling the rest of the image as stratiform, we use Parker and Johnson's (2000) definition of stratiform (20-40 dBZ) to identify the stratiform regions. We ignore regions with weaker reflectivity, and label regions with reflectivity above 40 dBZ as convective.

Another form of storm classification involves identifying storm cell areas and following them while gathering information about their strength. The Storm Cell Identification and Tracking (SCIT) algorithm does exactly this by finding reflectivity intensities that exceed thresholds for each of the seven tilts of the radar and then combining those tilts to find the cell areas (Johnson et al., 1998). Then the cell areas are compared across time steps to detect motion. The proposed methodology can also find cell areas, but since the main goal is to differentiate between storm types, we base our type on only one frame and do not track between frames. In addition, we do not match the storm areas between tilts since this technique requires data from only one level.

Our approach is most similar to Rigo and Llasat (2004). They combined aspects of the SCIT and SHY95 algorithms and used the combination as the basis for a structural classification system. Instead of using those storm areas to train an algorithm to automatically classify the storms in their dataset, they simply used the storm areas and statistics about them as guides when hand labeling each image. They also only assigned one storm type to each image even if more than one storm appeared whereas multiple storm types were often found and labeled within our dataset.

With this project we developed a structure-based classification system similar to that proposed by Rigo and Llasat. The propsed algorithm incorporates two machine learning techniques, k-means clustering and decision trees, to identify and classify storm areas.

The k-means clustering section of the project is derived from one used for image segmentation that had been applied to reflectivity (Lakshamanan, 2001; McQueen, 1967). Decision trees are structures that inductively sort a dataset by selecting attributes that lead to the most correct classification (Quinlan, 1986). They have the advantages of simplicity and an efficient implementation. In addition, when decision trees undergo the learning process, they can choose variables that help reach the decision and ignore the extraneous ones.

2. DATA AND METHODOLOGY

The data used for this project comes from both simulation and archived radar reflectivity from several storms in southwest Oklahoma. The simulations are generated by the Advanced Regional Prediction System (ARPS), a storm-scale model with numerical weather prediction and data assimilation features (Xue et al., 2001, 2002, 2003). We have over 250 simulations of mesoscale storms generated in a supercell regime (Rosendahl 2007). For this project, we examine the reflectivity at 4 km. The simulations used a 100 km by 100 km grid with 500 meter spacing. The reflectivity values of the simulated data tend to be higher than observed values of reflectivity because there is no attenuation or drop off in power with distance from the radar.

Our second source of data came from the Center for Collaborative Adaptive Sensing of the Atmosphere (CASA) IP1 network, a group of four small, X-band Doppler radars located in southwest Oklahoma (Brotzge et al, 2006). We mapped the reflectivity from each of the radars to a single 120 km by 120 km Cartesian grid with 500 m grid spacing to fit the image as closely as possible to the ARPS simulated data.

To identify individual storm regions, the program first divided a given reflectivity image into a specified number of clusters using the k-means clustering algorithm. To do this, the algorithm minimizes a Euclidean distance equation derived from the image segmentation algorithm of Lakshmanan (2001):

$$d_e = \lambda | r_m - r_p | + (1-\lambda)\sqrt{(x_m - x_p)^2 + (y_m - y_p)^2}$$

(Eq. 1)

In Eq. 1 λ weighs the differences in reflectivity versus Cartesian coordinates, r represents the reflectivity value in dBZ at a certain point, x and y are the coordinates of that point, m designates variables derived from the list of means, and p designates variables derived from a point in the reflectivity image. The first part of the equation seeks to find the distance of each point from the reflectivity means while the second part finds the distance between the selected point and the coordinates of the reflectivity means in the image. We chose a λ of .6 through empirical testing. K-

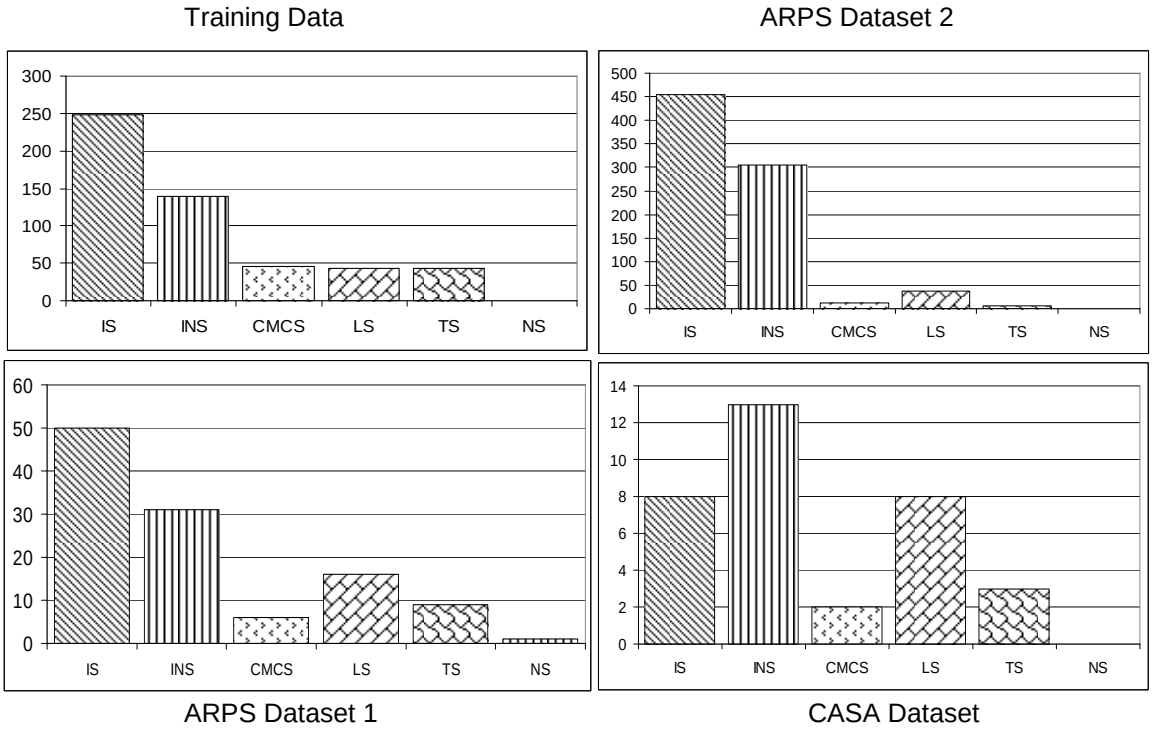

Fig. 1. The distributions of the four hand-labeled datasets used to train and test the decision trees.

means clustering uses this similarity metric to find geographically similar areas with similar reflectivity readings. We process the output of k-means clustering by breaking clusters that are not contiguous and by removing clusters whose area is less than 4km^2.

We divide the clusters into convective, stratiform, or a low reflectivity areas. If at least 70 percent of the cluster contains reflectivity between 20 and 40 dBZ, then it is considered stratiform. Otherwise, if less than 10 percent of the cluster contains reflectivity greater than 80 percent of the maximum reflectivity, then the cluster is considered a low reflectivity area. If the cluster fits neither of those categories, then it is considered convective.

Morphological	Reflectivity	Control
Eccentricity	Maximum	Area
Maj. Axis Len.	Minimum	Mean St. Dist.
Min. Axis Len.	Mean	
Orientation	Std. Deviation	
Equiv. Diam.	Range	

Table 1. Attributes used in decision trees.

The decision tree attributes are shown in Table 1. The morphological attributes come from fitting the storm region to an ellipse. These include the coordinates of the centroid, the length of the major and minor axes, the orientation of the ellipse in relation to the major axis and the horizontal, and the eccentricity of the ellipse. The reflectivity attributes include the maximum, minimum, mean, standard deviation, and the range of the reflectivity. The final attribute is the mean stratiform distance. In order to calculate this value, the program finds the equation of the line along the major axis and sets it equal to the offset value that must be added to the line in order to have the point that is the mean of the cluster centroids go through the line. It uses Equation 2.

$$d_s = m_l(x_s - x_c) + (y_c - y_s) \quad \text{(Eq. 2)}$$

In Equation 2 m_l represents the slope of the line, the *s* values represent centroid coordinates of the stratiform cluster and *c* values represent the same for the convective clusters. High positive values correlate more with the mean stratiform sitting in front of the convective area, values near 0 indicate that the mean stratiform center lies almost along the line, and high negative values indicate the mean stratiform lies behind the line.

Clusters are labeled within a hierarchal classification system that combines types developed by Parker and Johnson (2000) and Rigo and Llasat (2004). At the highest level, convective areas are divided into cell and linear mesoscale convective system (MCS). Within the

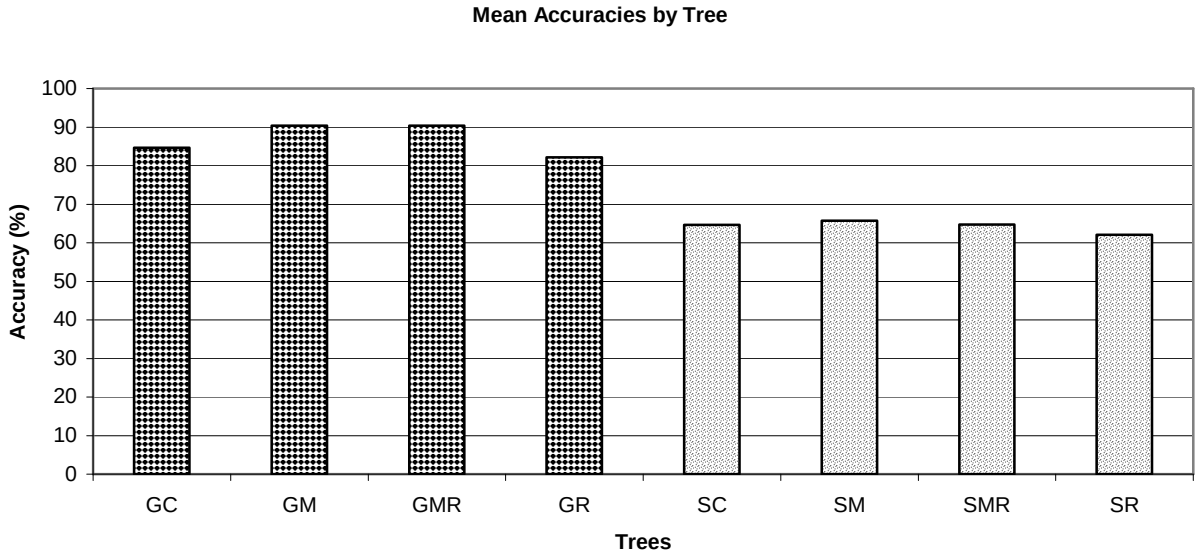

Fig. 2. This chart compares the accuracies of the eight different trees used in the study. The general trees (GC, GM, GMR, and GR) overall have a higher accuracy than the specific trees (SC, SM, SMR, and SR).

cell category, the storms are then subdivided into non-severe cell, severe cell, and circular MCS. In the linear MCS category, storms are divided into leading stratiform, trailing stratiform, and no/parallel stratiform.

The hand labeling interface provides the basic information required to visually classify storms. It displays an image of the reflectivity on the left side and a diagram of the convective cluster locations on the right side. From here when a clustered storm appears on the screen, we can select an individual cluster and then choose the appropriate classification for it from a pull down menu. Two meteorology students performed the hand labeling. Non-severe cells tended to be small areas of light to moderate reflectivity. Severe cells tend to be isolated areas of high reflectivity combined with other features indicating a powerful storm such as an overshooting top or a hook echo. Circular MCSs were generally clusters that contained multiple areas of high reflectivity intermixed with weaker reflectivity. Since the reflectivity image for the ARPS data is centered on the main storm rather than on a fixed point making the actual direction of storm motion impossible to determine, a stratiform area to the east of the storm is considered leading while one to the west is considered trailing.

To form the training and validation sets of the storms, we hand labeled five random time steps from each of the ARPS simulations. The decision trees were learned using the Waikato Environment for Knowledge Analysis (WEKA) version 3.5.6, developed by the University of Waikato in New Zealand (Witten and Frank, 2005) was used to generate the decision trees. WEKA is a suite of various machine learning and data mining algorithms. For the purposes of this project, we used the data in WEKA to generate eight decision trees based on different combinations of statistical values from the storm data. We generated decision trees for the general and specific storm types based on the morphological and reflectivity-based variables, just the morphological, just the reflectivity, and a control tree given only area and mean stratiform distance.

3. RESULTS

Figure 1 shows the distribution of the ARPS training data. Because of the regime used to generate the simulations, the storm types focus on isolated severe and isolated non-severe categories with 387 of the 519 storms falling into one of those two areas. The rest of the storms were almost evenly distributed across circular MCS, leading stratiform, and trailing stratiform. None of the storms in the training set were labeled as no stratiform. ARPS data set 1 contained a similar distribution of storms but in a much smaller number as it was randomly drawn from the same set as the training data. ARPS set 2 contained an even higher proportion of isolated severe and non-severe storms and very few of the other types. The CASA set featured a higher proportion of

isolated non-severe and leading stratiform storms, but it only held a total of 34 storms from three days.

WEKA learned eight decision trees based on different combinations of attributes. Trees designated as control (C) used only area and mean stratiform distance in their generation process. Morphological trees used the control attributes plus the morphological attributes derived from fitting the storm area to an ellipse. Reflectivity trees (R) used the control attributes reflectivity-based attributes. Morphological and reflectivity trees (MR) used all three classes of attributes.

We evaluated all the test sets on the eight different decision trees learned by WEKA. First we examined the overall accuracy of each tree (Fig. 2). The morphological and reflectivity tree (GMR) and the morphological (GM) tree had the highest mean accuracy at 90.39% while the tree that only used reflectivity attributes (GR) had the lowest mean accuracy at 82.17%. For the specific tree types, the one with the highest mean accuracy used the morphological attributes (SM) at 65.695% while the lowest came from the tree that used only reflectivity (SR, 62.06%).

To compare the performance of the trees on the different storm types we used area under the Receiver Operating Characteristic (AUC) curve (Bradley, 1997), which is a performance measure derived from plotting the probability of a true-positive result versus the probability of a false-positive result and finding the area under the resulting curve. The values range from 0 to 1 with values near 1 indicating a strong classification ability, values of .5 indicating that the classification system works no better than randomly choosing classes, and values of less than .5 indicating that the classification scheme is worse than random. For the general type trees (Fig. 3), the cell morphological and linear MCS morphological trees performed the best in both cells and linear systems with a mean AUC value of .912. This is averaged over all test sets. The lowest mean AUC for the general types, .836, came from the cell reflectivity and linear MCS reflectivity trees. When the AUC for each type was compared across the different datasets, there was very little difference in the AUC values.

Within the isolated severe cell type (Fig. 4), control, morphological, and reflectivity ranged from .802 to .782 while morphological and reflectivity combined had a lower value of .743. All the isolated non-severe storms performed between .85 and .89. For the circular MCS storms, three of the four trees received an AUC in the low .8s, but the morphological value lay at .689, much lower than the rest. Across each linear type (Fig. 5), there was little variance between each tree in each set although leading stratiform performed better than trailing stratiform.

4. DISCUSSION

The trees containing morphological attributes perform just as well as the trees containing both morphological and reflectivity attributes and better

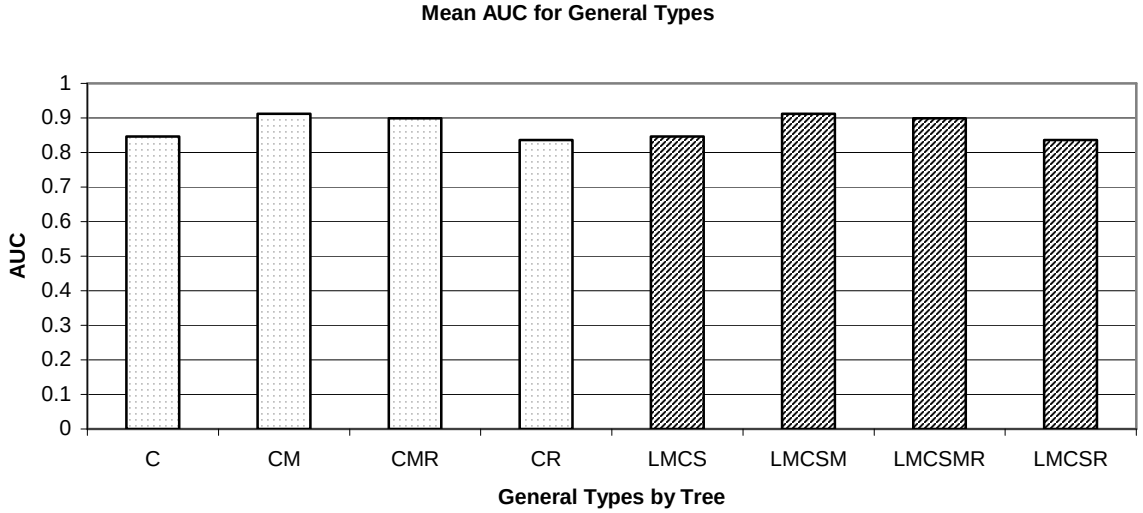

Fig. 3. This chart compares the area under a receiver operating curve (AUC), a measure of the classification system's reliability. Values near 1 indicate strong classification ability, at .5 indicate the same ability as a random classifier, and less than .5 indicate ability worse than random.

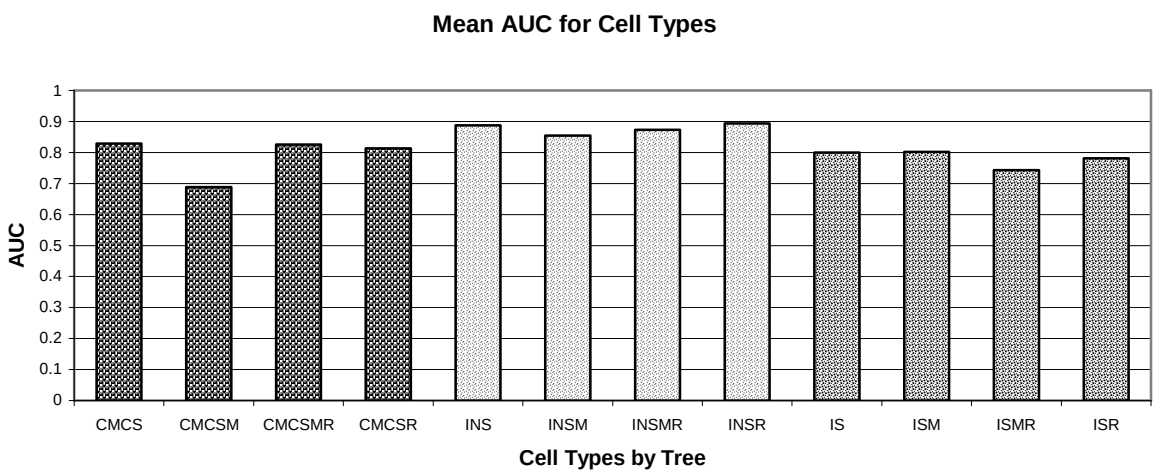

Fig. 4. This chart compares AUC among cell types. Little variance exists across the different tree types in each cell type except for the circular MCS (CMCS). The circular MCS morphological tree has a noticeable drop in performance in comparison to the other trees.

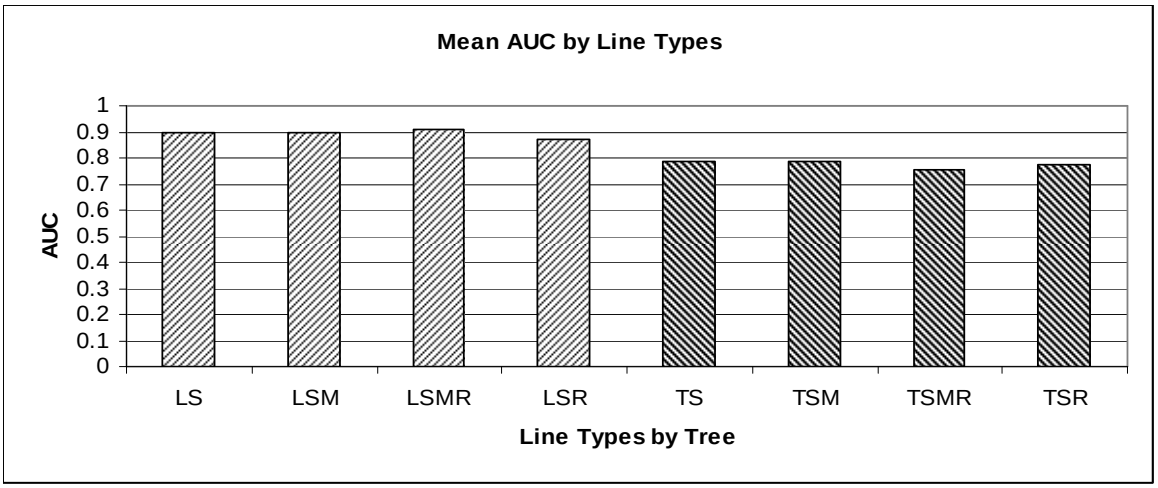

Fig. 5. This chart compares the mean AUC among the two linear MCS types found in the data. There is little variance within each type, indicating that one of the control variables most affects the tree performance on that type.

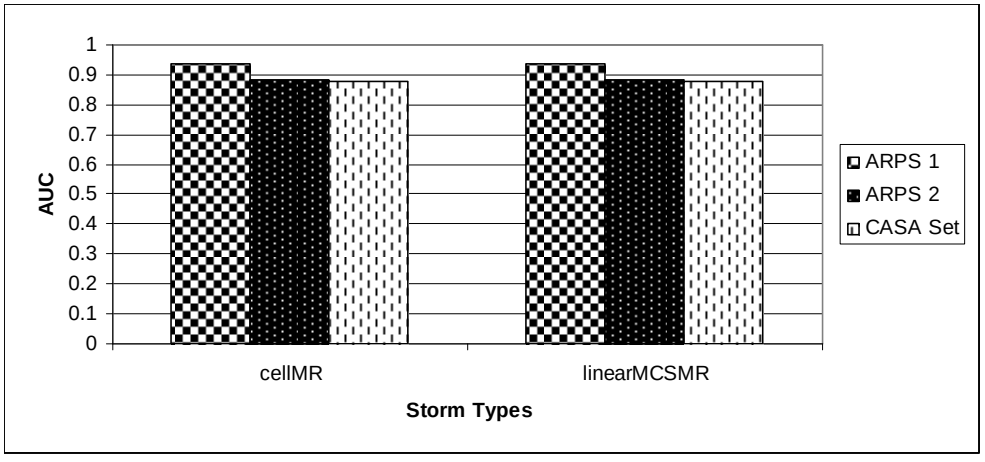

Fig. 6. This chart compares the AUC among the datasets for the general types.

perform just as well as the trees containing both morphological and reflectivity attributes and better than the ones only containing reflectivity attributes. The mean AUC for cells and linear systems does not vary significantly across the three datasets (Fig. 6). The fact that the tree learned using simulated data performed well on actual radar data (the CASA IP1 data) is critical as it means the trees were general across both simulated and observed data.

When analyzing the performance of the specific type trees, the influence of the reflectivity variables becomes more apparent. This is especially true for the circular MCS morphological tree, which has a much lower AUC than the circular MCS morphological and reflectivity and circular MCS reflectivity trees, indicating one of the reflectivity-based variables is the determining factor for this storm type. The same relationship exists for isolated non-severe. With isolated severe, however, the dip occurs with the MR tree. For the two linear types, the ROC area varies little across each type, indicating that the attribute for determining line type was one of the control attributes, area or mean stratiform distance.

5. CONCLUSIONS

We have found that decision trees are a viable method for automatically determining storm type. The trees that distinguished between cells and lines had a high AUC and accuracy across all datasets, indicating strong performance overall. The more specific trees experienced decreasing performance across datasets, which is an area for future work. Even though we learned the trees on simulated data, they were still able to classify real world data (the CASA data) with a high degree of accuracy in the case of the general type tree. In addition, because the decision trees are selective and human readable, we could determine that the morphological attributes were most critical to successful classification.

ACKNOWLEDGEMENTS

This work is supported in part by the Engineering Research Centers Program of the National Science Foundation under NSF award number 0313747. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation. We would like to thank Dr. Keith Brewster for providing the programs and assistance for extracting the CASA reflectivity data. In addition, we offer our thanks to Nathan Hiers for his help with labeling the ARPS 2 data.

REFERENCES

Anagnostou, E., 2004: A convective/stratiform precipitation classification algorithm for volume scanning weather radar observations, *Meteorological Applications*, **11**, 291-300.

Baldwin, M.E., J.S. Kain, and S. Lakshmivarahan, 2005: Development of an Automated Classification Procedure for Rainfall Systems. *Mon. Wea. Rev.*, **133**, 844–862.

Bradley, A. P., 1997: The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition,* **30**, 1145–1159.

Biggerstaff, M.I., and S.A. Listemaa, 2000: An Improved Scheme for Convective/Stratiform Echo Classification Using Radar Reflectivity. *J. Appl. Meteor.*, **39**, 2129–2150.

Brotzge, J., Droegemeier, K. K., and McLaughlin, D. J., 2006: Collaborative adaptive sensing of the atmosphere (CASA): New radar system for improving analysis and forecasting of surface weather conditions. *Journal of the Transportation Research Board*, (1948):145–151.

Doswell, C. A., III, H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting,* **11,** 560–581.

Johnson, J.T., P.L. MacKeen, A. Witt, E.D. Mitchell, G.J. Stumpf, M.D. Eilts, and K.W. Thomas, 1998: The Storm Cell Identification and Tracking Algorithm: An Enhanced WSR-88D Algorithm. *Wea. Forecasting*, **13**, 263–276.

Lakshamanan, V., 2001: A Hierarchal, Multiscale Texture Segmentation Algorithm for Real-World Scenes, Ph.D. dissertation, University of Oklahoma, 115 pp.

Maddox, R. A., D. M. Rodgers, and K. W. Howard, 1982: *Mesoscale convective complexes over the United States during 1981- annual summary, Mon. Wea. Rev.*, **110**, 1501–1514.

McQueen, J.B., 1967: Some Methods of Classification and Analysis of Multivariate Observations, *Proc. 5th Berkeley Symp. Mathemat. Statist. Probability*, **1**, 281–296.

Parker, M.D., and R.H. Johnson, 2000: Organizational Modes of Midlatitude Mesoscale Convective Systems. *Mon. Wea. Rev.*, **128**, 3413–3436.

Quinlan, J.R., 1986: Induction of decision trees, *Machine Learning*, **1**, 81-106.

Rigo, T. and M. C. Llasat, 2004: A methodology for the classification of convective structures

using meteorological radar: Application to heavy rainfall events on the Mediterranean coast of the Iberian Peninsula, *Natural Hazards and Earth System Sciences*, **4**, 59-68.

Rosendahl, D. H. (2007). Identifying Precursors to Strong Low-Level Rotation within Numerically Simulated Supercell Thunderstorms: A Data Mining Approach. Master's thesis, University of Oklahoma.

Steiner, M., R.A. Houze, and S.E. Yuter, 1995: Climatological Characterization of Three-Dimensional Storm Structure from Operational Radar and Rain Gauge Data. *J. Appl. Meteor.*, **34**, 1978–2007.

Witten, I., Frank, E., 2005: *Data Mining: Practical machine learning tools and techniques*. 2nd edition. Morgan Kaufmann, 525 pp.

Xue, M., Droegemeier, K. K., and Wong, V., 2000: The Advanced Regional Prediction System (ARPS) - a multiscale nonhydrostatic atmospheric simulation and prediction model. Part I: Model dynamics and verification, *Meteor. Atmos. Phys.*, **75**:161–193.

Xue, M., Droegemeier, K. K., Wong, V., Shapiro, A., Brewster, K., Carr, F., Weber, D., Liu, Y., and Wang, D., 2001: The Advanced Regional Prediction System (ARPS) - a multiscale nonhydrostatic atmospheric simulation and prediction tool. Part II: Model physics and applications, *Meteor. Atmos. Phys.*, **76**:134–165.

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, *Meteor. Atmos. Phys.*, **82**:139–170.