

RAINFALL CLASSIFICATION USING HISTOGRAM ANALYSIS: AN EXAMPLE OF DATA MINING IN METEOROLOGY

MICHAEL E. BALDWIN*

Cooperative Institute for Mesoscale
Meteorological Studies, University
of Oklahoma Norman, OK

also affiliated with: NOAA/NWS/Storm Prediction Center
and NOAA/National Severe Storms Laboratory

and

S. LAKSHMIVARAHAN

School of Computer Science,
University of Oklahoma
Norman, OK

ABSTRACT

The purpose of this work is to develop a robust automated classification technique to find significant and interesting features within spatial fields of meteorological data, for eventual use in a weather-related decision support system. To begin this analysis process, rainfall events are classified by analyzing the similarity of bulk statistical measures representing the distribution of rainfall values across a region of fixed size. The gamma distribution was selected to fit the observed distribution of rainfall since it is well suited for rainfall data. Due to the spatially correlated nature of rainfall, a robust method of parameter estimation of the distribution is required, therefore the generalized method of moments estimation technique was selected. Hierarchical cluster analysis is then performed using the parameters of the gamma distribution as attributes to classify the objects in the target data set, and those results are compared to a subjective classification of the rainfall patterns. The results show that this system successfully classified the cases in the target data set into convective and non-convective events with over 90% accuracy. However, further refinement of the classification was less successful and leaves room for future improvement. **Keywords:** data mining applications, parameter estimation

INTRODUCTION

Our overall goal is to develop a robust automated technique to classify significant and interesting features within a two-dimensional spatial field of meteorological data, such as observed or predicted rainfall. Ultimately, this classification system will be used in a weather-related decision support system. Analysis of such a complex data set can be made at several levels; similarity of the raw values of the variables at every point in space, 2-D image processing, spectral analysis, etc. As a first step in this multi-faceted analysis process, we choose to classify events by analyzing the similarity of bulk statistical measures representing the distribution of variable values across a region of fixed size. An initial target data set has been collected to test various data mining techniques. This data set consists of 1h accumulated rainfall analyses on a regular grid covering a 500km by 500km (approximately) region for 48 separate precipitation events occurring at

different times and locations across the United States. Each of these 48 events are considered “objects” for classification. The data set is relatively small and manageable but well-populated with interesting rainfall events that are desirable for classification. In order to reduce the dimensionality of the problem, bulk statistical parameters representing the distribution of rainfall amounts across the region are used as “attributes” for the classification. The gamma distribution was selected since it is well suited for rainfall data and has been widely used for rainfall histogram analysis in the meteorological literature. Due to the spatially correlated nature of rainfall, a robust method of parameter estimation of the gamma distribution is required, therefore we selected the generalized method of moments (GMM) estimation technique. Hierarchical cluster analysis is then performed using the parameters of the gamma distribution as attributes to classify the objects in the target data set, and those results are compared to a subjective classification of the rainfall patterns. The results show that this system successfully classified the cases in the target data set into convective and non-convective events with over 90% accuracy. Further refinement of the classification is an open problem and is left for future work.

Due to space constraints, we only provide a summary of this work. A more detailed report is provided by Baldwin and Lakshminarayanan (2002). The remainder of this paper will proceed as follows. The next section will outline the test data set. The section following that will describe the choice of the statistical model and parameter estimation technique, along with the classification algorithm. Analysis of results and concluding remarks will be provided in the last section of the paper.

TARGET DATA SET

To begin this work, a small target data set is established. The so-called “Stage IV” rainfall analysis (Baldwin and Mitchell 1998) produced at the National Centers for Environmental Prediction (NCEP, an agency of the U.S. National Weather Service responsible for creating and delivering environmental information and forecasts to both the public and private sectors) was obtained for the period covering late summer/early fall of 2000. The Stage IV analysis is a national mosaic of optimal estimates of hourly accumulated rainfall using radar and raingage data. The domain size was chosen to be fixed at 128 x 128 4km grid boxes, which is approximately 500km by 500km. A set of 48 cases was selected for inclusion in the target data set. The selection criteria was based upon the occurrence of “typical” rainfall patterns that often occur across the U.S. during the year. Each case was subjectively classified (by a NCEP meteorologist) into the set of event classes and sub-classes listed in Table 1. Convective precipitation events are produced by small-scale (wavelengths on the order 100km and smaller), convectively-driven atmospheric circulations. In the linear sub-class, the precipitation field is fairly consistent along a line, with a large variation in the direction normal to the line, such as a squall line. For the cellular sub-

Table 1: Hierarchy of rainfall classes (case numbers).

	Convective	Non-convective	
Linear (1-16)	Cellular (17-34)	Orographic (35-40)	Stratiform (41-48)

class, the precipitation field consists of nearly circular-shaped features. Non-convective precipitation events are produced by upward vertical motion resulting from large-scale (wavelengths on the order 1000km and larger) forcing mechanisms. In the stratiform subclass, the precipitation field shows little variation in any direction over a large area. For the orographic subclass, the precipitation field is strongly tied to the shape of the terrain field.

METHODOLOGY

There are a large number of potential choices of attributes that could describe the rainfall pattern over a region. An obvious choice is the amount of rainfall at every point in space obtained from a gridded analysis over the region of interest. Since the goal of this work is to identify the dominant type of rainfall “event” that is found within a region, such as storms oriented along a line or a cluster of cellular-type convection, the precise location of the maxima/minima is not of great importance. Therefore, a logical choice for the attributes might be some sort of bulk statistical measure of the overall distribution of rainfall across the region. To begin this work, the simplest choice of bulk statistical measures was selected, that is the parameters of a theoretical statistical distribution fitted to the histogram representing the observed distribution of rainfall amounts across the region of interest. The distribution of rainfall tends to be highly positively skewed. For example, heavy rainfall is a rare event, and when large amounts of rain do occur, such as typically found intermittently in some convective systems, the resulting distribution possesses a long “tail”. It is also common to see widespread light rain, such as typically found in non-convective systems, resulting in a distribution that is “humped” near a low amount of rainfall with little or no “tail”. These characteristics limit the choices of theoretical distributions as potential models for the observed distribution. For this work, we selected the gamma distribution since it is positively skewed and non-negative, provides a reasonable representation with only two parameters, and has been widely used in the meteorological literature for the analysis of precipitation data (e.g., Wilks 1990).

Rainfall data, like many meteorological variables, are spatially correlated. For this reason, a robust method of parameter estimation is desired that does not rely upon an assumption of independence, for example, the generalized method of moments (GMM, Hamilton 1994). An overview of GMM is provided in Baldwin and Lakshmivarahan (2002). GMM can allow for correlation in the data to affect the parameter estimation. GMM can be considered an extension to the more familiar method of moments technique for parameter estimation. In the method of moments technique, a set of equations are developed to cover the number of unknown parameters found in the model. In the case of the gamma distribution, there are two unknown parameters, α and β , therefore two equations relating these to known quantities are needed. Solving this system of equations provides an estimate of α and β , the resulting distribution will fit the observed mean and variance exactly, but higher-order moments are not taken into account. In some cases, it may be desirable for the parameters to provide a better fit to the observed skewness (related to the 3rd moment) or kurtosis (related to the 4th moment). The GMM technique

allows for this by adding higher-order moments to the equation set, resulting in a non-linear system of equations which can then be solved by least-squares methods. In this work, we tried several different combinations of moments (2-4) and values of the lag-correlation in the data ($q=0$ to 5) in estimating the gamma parameters. These estimates of α and β are then used in a classification algorithm in order to find groups of similar rainfall events. To our knowledge, this work is the first example of the use of GMM with rainfall data in the meteorological community.

There are a variety of data mining algorithms to choose from that have been developed to collect groups of objects with similar attributes. Since classification is the desired data mining task in this work, hierarchical cluster analysis has been selected as the primary classification tool for this work. Here, objects will be clustered where objects are defined as rainfall events over regions of fixed size. The goal of this hierarchical classification scheme is to first group the cases into convective/non-convective classes, then further refine these classes into linear/cellular for the convective class and stratiform/orographic for the non-convective class. The hierarchical cluster analysis method that is chosen for this work is Ward's method (Alhamed et al. 2002), which is based upon the fact that the total scatter (or variance) of all of the objects is constant and can be partitioned into the between-cluster scatter and the within-cluster scatter. Ward's method has been found to produce good results for meteorological data in previous research (Alhamed et al. 2002).

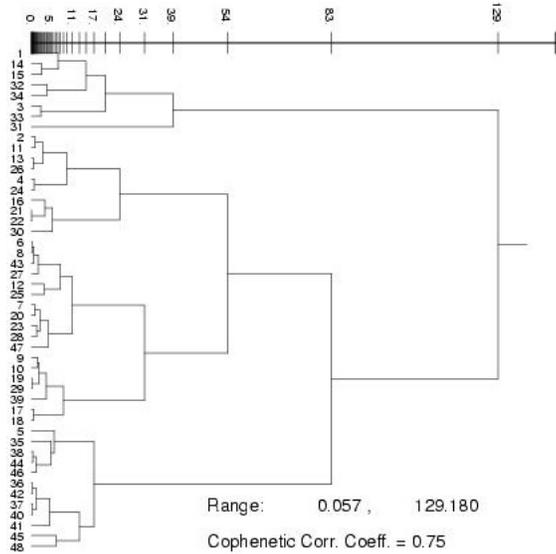


Figure 1: Dendrogram produced by Ward's method using 2-moment GMM, no lag-correlation.

ANALYSIS OF RESULTS AND CONCLUSIONS

Figure 1 shows a sample dendrogram of results from the Ward's method on the target data set for the 48 cases using α, β estimated by GMM using 2 moments (first and second) and no assumed lag-correlation. In the dendrogram, there appears to be four main clusters which are separated at the square error breakpoint (on the x-axis) of ~ 40 . The cases found within these four clusters are listed in Table 2. This result shows the CA successfully produces clusters whose members fall into the subjectively determined convective/non-convective classes. For example, clusters 1 and 2 are unanimously populated by convective-type events (both linear and cellular). Cluster 3 is dominated by convective events, with 3 (out of 18) exceptions (cases 39, 42, and 47). Cluster 4 is dominated by non-convective events, with 1 (of 12) exception (case 5). Overall, there are only 4 out of 48 "mis-classified" events, resulting in a 92% classification accuracy.

These results were similar to those found with three and four moments, and by increasing the lag-correlation value from zero to five, as summarized in Table 3. The classification shows some sensitivity to the choice of moments used in the parameter estimation. However, it does not appear to be sensitive to the choice of lag-correlation value used, even though the estimated α and β values varied as the assumed lag-correlation (q) changed (not shown). For each set of moments, the cluster analysis produced identical members for all values of q from 0 to 5. Among these experiments, the 3-moment (first, second, and third) GMM estimation produced the best classification when validated against the classes determined subjectively by a NCEP meteorologist.

Table 2: Cluster membership for the 2-moment, no assumed correlation experiment.

Cluster	Cases
1 (8)	1, 3, 14, 15, 31, 32, 33, 34
2 (10)	2, 4, 11, 13, 16, 21, 22, 24, 26, 30
3 (18)	6, 7, 8, 9, 10, 12, 17, 18, 19, 20, 23, 25, 27, 28, 29, 39, 42, 47
4 (12)	5, 35, 36, 37, 38, 40, 41, 42, 44, 45, 46, 48

Now we examine how well the cluster analysis classifies the cases into the four sub-classes (linear, cellular, stratiform, orographic). Returning to the 2-moment, $q=0$ experiment (Table 2), cluster 1 contains four cases that were subjectively classified as linear and four that were subjectively classified as cellular precipitation events. Cluster 2 is also evenly split among the linear and cellular precipitation events with five cases from each. Cluster 3 contains six linear events, nine cellular events, one orographic, and two stratiform events. Cluster 4 contains mainly stratiform (6) and orographic (5) events, with one linear event included. These results show that the CA did not produce clusters with a clear preference for a particular sub-class in this experiment. These results were similar to those found with three and four moments, and by increasing the lag-correlation value from zero to five, with some variation.

The hierarchical clustering algorithm successfully separated the cases into convective and non-convective classes. However, looking at the next level of classification hierarchy,

the four main clusters did not match the four sub-classes (linear, cellular, stratiform, orographic) very well. This should not be surprising, since two parameters (α, β) should be able to discriminate between two classes (convective, non-convective) quite well, but have some difficulty in further refining the classification. It is reasonable to expect that additional discriminants will be needed in order to increase the degrees of freedom and allow the classification system to identify finer and more specific classes of events. This sets the stage for future work where we will use; cluster analysis to classify events based upon similarity of the raw values at each point in space, principal component analysis to transform the data, image processing techniques to refine the selection of attributes, etc. The choice of attributes is obviously critical, attributes based upon some measure of the spatial variability and intermittence (Harris et al 2001) of the fields could help in refining the classification to uncover the sub-classes of convective (linear, cellular) and non-convective (orographic, stratiform) precipitation events.

Table 3: Results of cluster analysis into convective/non-convective classes

Experiments	Mis-classified cases (percent correct)
2-moments; $q=0, 1, 2, 3, 4, 5$	5, 39, 43, 47 (92%)
3-moments; $q=0, 1, 2, 3, 4, 5$	5, 43, 47 (94%)
4-moments; $q=0, 1, 2, 3, 4, 5$	5, 9, 10, 43, 47 (90%)

ACKNOWLEDGEMENTS

Professor Daniel Wilks, Cornell University, kindly provided the software for maximum likelihood estimation of the parameters of the gamma distribution. Ying Lin of the Environmental Modeling Center provided archived rainfall analysis data. Ahmed Alhamed provided cluster analysis software. Other software for solving non-linear least squares and matrix inversion was obtained from the Netlib.org repository.

REFERENCES

- Alhamed, A., S. Lakshmvirahan, and D. J. Stensrud, 2002: Cluster analysis of multimodel ensemble data from SAMEX. *Mon. Wea. Rev.*, **130**, 226-256.
- Baldwin, M. E., and S. Lakshmvirahan, 2002: Rainfall classification using histogram analysis: An example of data mining in meteorology. Technical Report, School of Computer Science, University of Oklahoma, Norman, OK.
- Baldwin, M. E., and K. E. Mitchell, 1998: Progress on the NCEP hourly multi-sensor U. S. precipitation analysis for operations and GCIP research. Preprints, 2nd Symposium on Integrated Observing Systems, 78th AMS Annual Meeting, January 11-16, 1998, Phoenix, Arizona, 10-11.
- Hamilton, J. D., 1994: *Time Series Analysis*. Princeton University Press, 799pp.
- Harris, D., E. Foufoula-Georgiou, K. K. Droegemeier and J. J. Levit, 2001: Mutiscale statistical properties of a high-resolution precipitation forecast. *Journal of Hydrometeorology*, **2**, 406-418.
- Hosking, J. G., and C. D. Stow, 1987: Ground-based, high resolution measurements of the spatial and temporal distribution of rainfall. *J. Climate Appl. Meteor.*, **26**, 1530-1539.
- Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Clim.*, **3**, 1495-1501.