Using Bayesian Networks to Estimating Rainfall Distribution Given Polarimetric Radar Data

Jose Picado, Vahid Ghadakchi, Zahra Iman

1 Introduction

Measuring the amount of rainfall on a specific field is an important issue in agriculture. The basic way to measure the amount of rainfall is to plant rain gauges on the ground and use them to carry out the measurement. However it is not possible to plant gauges on every field. Therefore remote sensing instruments such as radars are applied to measure the amount of rainfall. The output of radars do not exactly match the measurements of the gauges. The output of radars are corrected using the nearby gauges and a single distribution is provided to users. In this project, we will address the problem of finding the probability distribution of rainfall using Bayesian networks [1].

The obtained distribution using Bayesian network highly depends on the structure of the network and the relation of variables in the network. We explore two approaches to build the structure of a Bayesian network. First, we manually design the structure of the network based on domain knowledge. Domain knowledge is obtained by studying the 18 features provided in the dataset and figuring out the dependencies that hold between these variables. Second, we learn the structure of the network using learning algorithms. This paper is structured as follows: In Section 2 we will provide a brief review of studies on different features of the dataset. In Section 3 we present our methodology, where we introduce the data, analyze the dependencies between features, and present the Bayesian network structures. Section 4 contains the results of predicting the probabilities using both the manually designed Bayesian network and the learned Bayesian network, and a comparison to two baseline methods.

2 Related Work

Weather radars have been used for locating precipitation and estimating different types of weather situations for many years. By introduction of Polarimetric Doppler radars, new studies have been conducted to provide more accurate estimation of rainfall rate. The most basic method for rainfall rate estimation was based on the reflectivity measurement provided in equation 1.

[1] https://www.kaggle.com/c/how-much-did-it-rain

In (Zrnic 1999) Zrnic provides the information about the relationship of differential phase (K_{DP}) and rainfall rate. It is shown that this relation is almost linear which is shown in equation 2. This method outperforms the conventional rainfall rate estimation method based on reflectivity in cases that hail contamination, partial radar beam blockage, attenuation in rain exists. But, it is mentioned that differential phase is not sensitive to drop size distribution variations, presence of dry and tumbling hail.

On the other hand, the estimates of rainfall rates had a tendency to be noisier when the sizes of drops were smaller. Seliga et al (Seliga and Bringi 1976) introduced the differential reflectivity parameter (Z_{DR}), which is defined as the ratio of radar reflectivities at horizontal and vertical polarization. This parameter is sensitive to size of drop and therefore rainfall rate estimates based on this parameter joint with radar reflectivity provide more accurate estimations with smaller errors. But, there are some biases introduced into Z_{DR} measurements due to mismatches in radar beam patterns at horizontal and vertical polarization and sidelobes.

Gorgucci and Scarchilli (Gorgucci, Scarchilli, and Chandrasekar 1994) suggest the formula presented in equation 3 to estimate rainfall rate.This method combines the differential reflectivity (Z_{DR}) information with differential phase (K_{DP}).

The fourth type shown in equation 4 combines the differential reflectivity with radar horizontal reflectivity (Z) in order to estimate rainfall rate (Bringi and Chandrasekar 2001).

In the following formulations, Z shows the radar reflectivity at horizontal polarization, K_{DP} shows the differential phase, Z_{DR} shows the differential reflectivity.

$$R(Z) = 0.017Z^{0.714} \tag{1}$$

$$dataR(K_{DP}) = 42.8|K_{DP}|^{0.802} sign(K_{DP}) \tag{2}$$

$$R(K_{DP}, Z_{DR}) = 51.0K_{DP}^{0.968} Z_{DR}^{-0.462} \tag{3}$$

$$R(Z, Z_{DR}) = 0.0067Z^{0.93} Z_{DR}^{-3.43} \tag{4}$$

Ryzhkov in (Ryzhkov, Schuur, and Zrnic 2002) argues that the last two formulations are not robust at low values of Z_{DR}. Therefore they suggest the following equations for estimation of rainfall rates.

$$\left\{ \begin{array}{cc} R(K_{DP}, Z_{DR}) = 53.7 K_{DP}^{0.91} Z_{DR}^{-0.421} & Z_{DR} > 0.5 \\ R(K_{DP}, Z_{DR}) = 70.0 K_{DP}^{0.878} 10^{-0.131 Z_{DR}} & Z_{DR} < 0.5 \end{array} \right\} \quad (5)$$

$$\left\{ \begin{array}{cc} R(Z, Z_{DR}) = 0.0064 Z^{0.824} Z_{DR}^{-0.654} & Z_{DR} > 0.5 \\ R(Z, Z_{DR}) = 0.014 Z^{0.813} 10^{-0.266 Z_{DR}} & Z_{DR} < 0.5 \end{array} \right\} \quad (6)$$

The goal of this project is to predict the rainfall rates based on the polarimetric radar measurements and other features extracted from these readings. In this regard, we would take advantage of having more features. For example, the rain rates extracted from the afformentioned methods would in addition Hydrometeor type and correlation coefficient features would be used as the input features to our method. Methodology

2.1 Features and Dependencies

In this Section, we describe each feature present in the radar measurements. This information is used to manually build a Bayes network structure, as explained later.

- TimeToEnd: This features shows that the specified radar observation time in terms of minutes left to one hour observation. This feature is dependent on measurements features because the time to end feature shows the different times of measuring the values which is shown in the measurement feature after averaging.
- DistanceToRadar: This feature shows the distance from radar to gauge. Each value in this column maps to each time value in the TimeToEnd column. Therefore there can be multiple readings in 1 hour. In the same sense as time to end, distance to radar feature had different measurements of different times and therefore is related to the measurement column.
- Composite: This feature shows maximum reflectivity in vertical column above gauge i.e. the maximum dBZ reflectivity from any of the reflectivity angles. Reflectivity shows the precipitation intensity at that specific angle above the horizon. In the Composite, the highest intensities amongst those available in the different angles above each point of the image, will be considered.
- HybridScan: Reflectivity in elevation scan closest to ground. The objective is to utilize reflectivity measurements from as close to 1-km altitude above radar level as possible while minimizing the likelihood of ground clutter and data loss due to terrain blockages. This feature is dependent on composite value. It is mentioned in (com 2004) that NEXRAD scans in several angles where the radar makes a 360-degree horizontal sweep with the radar antenna tilted at the given angle above the horizontal, then changes the elevation angle, and completes another 360-degree sweep, and so on. Therefore the hybrid scan as being the reflectivity closest to ground is dependent on this scan.
- HydrometeorType: One of nine categories in NSSL HCA: no echo, moderate rain, heavy rain, rain/hail, big drops, AP, Birds, Unknown, dry snow, wet snow, ice crystals, graupel. Based on the automatic classification of hydrometeor types algorithm introduced in (Zrnic et al. 2001), the algorithm takes as input reflectivity, differential reflectivity (Z_{DR}), specific differential phase (K_{DP}), cross-correlation coefficient ($RhoHV$). Therefore, the HydrometeorType is dependent on these features in the proposed Bayesian Network structures.
- Reflectivity (in dBZ): A traditional non-polarised radar will measure only the reflectivity at each gate which is the same as horizontal reflectivity in Polarimetric Doppler radars.
- K_{DP} : Differential phase. The specific differential phase is a comparison of the returned phase difference between the horizontal and vertical pulses. This change in phase is caused by the difference in the number of wave cycles (or wavelengths) along the propagation path for horizontal and vertically polarized waves.
- $RhoHV$: Cross-correlation coefficient. A statistical correlation between the reflected horizontal and vertical power returns. It is a good indicator of regions where there is a mixture of precipitation types, such as rain and snow. High values, near one, indicate homogeneous precipitation types, while lower values indicate regions of mixed precipitation types, such as rain and snow, or hail.
- Z_{DR} : Differential reflectivity in dB: The differential reflectivity is the ratio of the reflected vertical and horizontal power returns as Z_V/Z_H. Among other things, it is a good indicator of drop shape and drop shape is a good estimate of average drop size.
- RR1: Rain rate from HCA-based algorithm. Based on (Laboratory 2007), HCA algorithm is used to remove non meteorological and therefore improve rainfall rate estimates. Therefore, RR1 is dependent on HydrometeorType, Z_{DR}, $RhoHV$ and K_{DP}.
- RR2: Rain rate from Z_{DR} -based algorithm. Rainfall rate estimation based on Z_{DR} is shown in equation 3. Therefore, this feature is dependent on Z_{DR}.
- RR3: Rain rate from K_{DP} -based algorithm. Rainfall rate estimation based on K_{DP} is shown in equation 2. Therefore, this feature is dependent on K_{DP}.
- Velocity: Doppler velocity. It gives only the radial variation of distance versus time between the radar and the target. The phase between pulse pairs can vary from $-$ and $+$, so the unambiguous Doppler velocity range is $V_max = /4t$
- LogWaterVolume: How much of radar pixel is filled with water droplets?
- MassWeightedMean: Mean drop size in mm. Based on (Bringi and Chandrasekar 2001), it is known that Z_{DR} and K_{DP} provide information about shape and size of rain drops. Therefore, this feature is dependent on these two radar measurements.
- MassWeightedSD: Standard deviation of drop size. For the same reason provided for MassWeightedMean fea-

ture, the MassWeightedSD is also dependent on Z_{DR} and K_{DP}.

- RadarQualityIndex: A value from 0 (bad data) to 1 (good data). This index is computed based on the Quality Control algorithm introduced in (Lakshmanan et al. 2007). They use a neural network to combine multiple features into a single discriminator that can distinguish between good and bad echoes. This algorithm operates on six moments available from polarimetric radar: Reflectivity (Z), Velocity (V), Correlation Coefficient ($RhoHV$), Differential Reflectivity (Z_{DR}) and Differential Phase (K_{DP}). Therefore, this RadarQualityIndex is dependent on these features.
- ReflectivityQC: Quality-controlled reflectivity. This feature is dependent on RadarQualityIndex and Reflectivity.
- Expected: The actual amount of rain reported by the rain gauge for that hour. This feature is dependent on Composite, QualityControlIndex, MassWeightedSD, MassWeightedMean, RR1, RR2, RR3, LogWaterVolume, MassWeightedMean, Velocity, DistanceToRadar and TimeToEnd.

2.2 Building the Bayesian Network Structure Manually

Based on the features and studied dependencies between them provided in Section 3.1, we propose the Bayesian network structure presented in Figure 1.

2.3 Learning the Bayesian Network Structure

Besides designing the Bayesian network structure manually, we also employed structure learning algorithms to learn the structure based on the data. We tried several structure learning algorithms that employ a Bayesian metric. The Bayesian metric assumes Dirichlet priors and is given by (Bouckaert 2008):

$$Q_{Bayes}(G,D) = P(G)\prod_{i=0}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})}$$

where r_i is the cardinality of variable x_i, q_i is the cardinality of the parent set of x_i in structure G, n is the number of random variables, N_{ijk} is the number of records in the dataset D for which $Pa(x_i)$ takes the jth value and for which x_i takes the kth value, and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. $P(G)$ is the prior on the network structure and $\Gamma(.)$ the gamma function. N'_{ij} and N'_{ijk} represent choices of priors on counts restricted by $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$.

We explored greedy algorithms such as Hill Climbing (Buntine 1996) and K2 (Cooper and Herskovits 1992). However we found that, for our domain, these algorithms were simply learning a network structure where all features are conditionally independent given the class (Naïve Bayes). Then, we tried Simulated Annealing (Bouckaert 1995) as an alternative to greedy approaches. We found this approach to give better results, which are presented in the next Section.

3 Results and Discussion

3.1 Data

We employed the dataset provided by (Lakshmanan 2015). This dataset contains measurements of Polarimetric radars. It consists of NEXRAD and MADIS data collected the first 8 days of April to November 2013 over Midwestern corn-growing states. Polarimetric radars give more accurate readings than conventional radars, as they can infer the drop size and therefore improve the estimate of rainfalls. Time and location information have been censored, and the data is not ordered by time or place. The test data consists of data from the same radars and gauges over the same months but in 2014.

For our experiments, we employed a subset of the training dataset for efficiency reasons. We randomly selected 10,000 training examples from the entire dataset, and used this to train all our models. We used the full testing set, which contains 630451 instances, as we needed to make predictions for all the instances in order to submit our results to the kaggle competition.

The data obtained from Polarimetric radars is feature-based. However some instances contain multiple readings for the same feature. That is, some features have composite values. To overcome this issue, we aggregated the composite values into one single value. The aggregate function that we used for most features was the mean function. Only for the TimeToEnd feature, we applied the following function:

$$V = \frac{max(time) - min(time) + 6}{60}$$

The number 6 assumes that readings are done in 6-minute timesteps. The number 60 works as a normalization constant.

Because we are aggregating composite values into single values, we may be losing some information. However, we believe that multiple readings (composite values) should be more accurate. Therefore, we added a feature that indicates the number of readings for the instance. This way, instances with multiple readings may have higher weights.

3.2 Experiments

We have used Weka (Hall et al. 2009), a popular set of machine learning tools implemented in Java, to carry out the experiments. As a benchmark, we have used Logistic Regression to predict the probability distribution of rain amount. We have also included the results of using Naïve Bayes and applying simple estimation of conditional probabilites. Simple estimation is based on the following formula (Bouckaert 2008):

$$P(x_i = k | Pa(x_i) = j) = \frac{N_{ijk} + \alpha}{N_{ij} + M \times \alpha}$$

where N_{ijk} is the number of instances where $Pa(x_i)$ takes the jth value and x_i takes the kth value, and N_{ij} is the

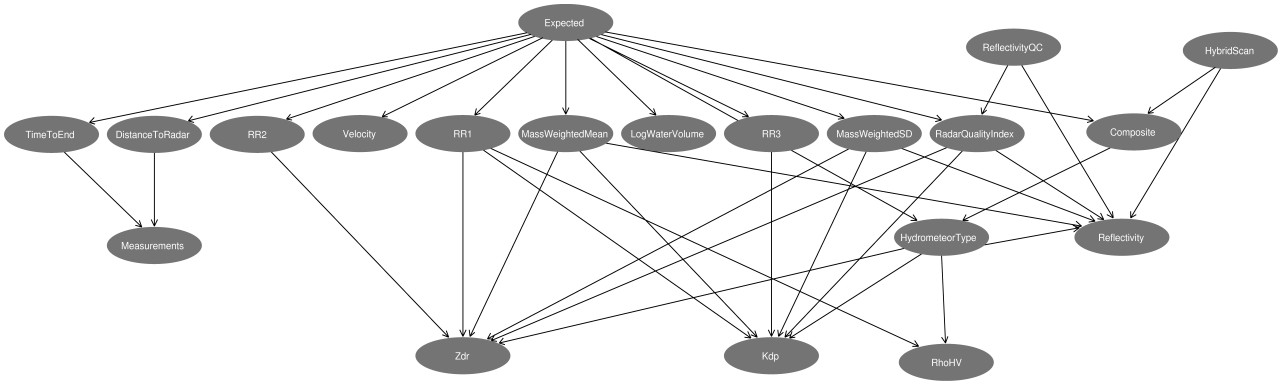

Figure 1: Proposed Bayesian network structure based on the study on features and their dependencies.

Algorithm	CRP
Logistic Regression	0.01096134
Naïve Bayes	0.01452578
Bayes Net - Manually Designed Structure	0.00965175
Bayes Net - Simulated Annealing	0.00864957

Table 1: Continuous Rank Probability (CRP) of different algorithms.

number of instances where $Pa(x_i)$ takes the jth. Parameter $\alpha = 0.5$ is used for smoothing. Setting it to zero will lead to maximum likelihood estimation. M is the size of data-set.

The results of employing different algorithms and approaches are presented in Table 4.2. Different approaches are compared based on Continuous Ranked Probability Score:

$$C = \frac{1}{70N} \sum_{N} \sum_{n=0} 69(P(Y \leq n) - H(n - z))^2$$

N is the size of test data set, z is the actual label, H is unit step function and P is the cdf for the amount of rain. C represents the distance between forecasted cdf and the real cdf and our goal is to minimize the value of C.

Greedy structure learning algorithms, such as Hill-Climbing (Buntine 1996) and K2 (Cooper and Herskovits 1992), generated a structure similar to Naïve Bayes. So they did not improve the final results. On the other hand, the network with the manually designed structure as explained in Section 3 takes into account the dependencies between different variables. Applying simple parameter estimation on this structure leads to better results comparing to our baselines, Logistic Regression and Naïve Bayes, and greedy structure learning algorithms. Using structure learning with a non-greedy algorithm, such as simulated annealing, showed to be very helpful, as it got the best results.

Results show that employing domain expert to build the structure of a network can be useful. This approach got better results than traditional learning algorithms, and even better results than Bayesian networks with structures learned in a greedy way. On the other hand, non-greedy structure learning algorithms proved to even get better results. Therefore, this approach is certainly the best approach when enough data is available and time is not an issue.

4 Conclusion

In this paper we explored the use of Bayesian networks to find the distribution of rainfall given measurements from Polarimetric radars. We employed two different approach for constructing the structure of the Bayesian network. First, we manually designed the structure based on domain knowledge. Doing this required an extensive study of the features present in the dataset and the dependencies between them. Second, we applied a structure learning algorithm to learn the structure automatically from data. Results showed that the Bayes network with the structure learned from data performed best. The Bayes network with manually designed structure did not do as well, but it was still better than our baselines.

As a future work, it would be interesting to have a domain expert study the features and provide the dependencies between them. It would also be interesting to try undirected models, such as Markov networks. This is because many times dependencies exist between features, but it is not clear what is the direction of the dependency. Furthermore, it would be interesting to try different aggregation functions for features, based on their nature. We employed the mean function for all features. However, other functions may be more appropriate for some features. Finally, learning both structure and parameters with the full dataset, instead of a subset, would be interesting. The dataset contains 70 labels, therefore a bigger dataset would give a better picture of the distribution of the labels.

References

Bouckaert, R. 1995. *Bayesian Belief Networks: from Construction to Inference*. Ph.D. Dissertation, Utrecht, Netherlands.

Bouckaert, R. 2008. Bayesian network classifiers in weka for version 3-5-7.

Bringi, V., and Chandrasekar, V. 2001. *Polarimetric Doppler weather radar: principles and applications*. Cambridge University Press.

Buntine, W. 1996. A guide to the literature on learning probabilistic networks from data. *Knowledge and Data Engineering, IEEE Transactions on* 8(2):195–210.

2004. Oklahoma climatological survey training materials.

Cooper, G. F., and Herskovits, E. 1992. A bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* 9(4):309–347.

Gorgucci, E.; Scarchilli, G.; and Chandrasekar, V. 1994. A robust estimator of rainfall rate using differential reflectivity. *Journal of Atmospheric and Oceanic Technology* 11:586592.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter* 11(1):10–18.

Laboratory, G. R. 2007. Rainfall estimation with dual-pol algorithms based on the results of new kdp processing.

Lakshmanan, V.; Fritz, A.; Smith, T.; Hondl, K.; and Stumpf, G. J. 2007. An automated technique to quality control radar reflectivity data. *J. Applied Meteorology* 46(3):288–305.

Lakshmanan, V. A. 2015. Probabilistic estimate of hourly rainfall from radar. *13th Conference on Artificial Intelligence*.

Ryzhkov, A.; Schuur, T.; and Zrnic, D. 2002. Testing a polarimetric rainfall algorithm and comparison with a dense network of rain gauges. *Hydrological Resources on Hydrological Applications of Weather Radar, Kyoto, Japan* 159–164.

Seliga, T., and Bringi, V. 1976. Potential use of radar differential reflectivity measurements at orthogonal polarizations for measuring precipitation. *Journal of Applied Meteorology* 15(1):69–76.

Zrnic, D. S.; Ryzhkov, A.; Straka, J.; Liu, Y.; and Vivekanandan, J. 2001. Testing a procedure for automatic classification of hydrometeor types. *Journal of Atmospheric and Oceanic Technology* 18(6):892–913.

Zrnic, D.S., A. R. 1999. Polarimetry for weather surveillance radars. *Bull. Amer. Meteor. Soc* 389 406.