Predicting Rainfall with Polarimetric Radar Data

Nitin Kamra
Department of Computer Science
University of Southern California
Los Angeles, CA 90007
nkamra@usc.edu

James Preiss
Department of Computer Science
University of Southern California
Los Angeles, CA 90007
japreiss@usc.edu

Abstract

We explore a set of polarimetric radar data and rain gauge readings collected in the Midwestern US over several months, while aiming to improve existing rainfall prediction techniques with various supervised learning algorithms.[1] We explore preprocessing techniques and evaluate Linear Regression and Feedforward Neural Networks models of summarized data against the Marshall-Palmer baseline.

1 Introduction

This work focuses on predicting hourly rainfall using measured data from polarimetric radars. Rainfall is a highly distributed phenomenon, and hence very hard to measure. The amount of rainfall is conventionally measured with rain gauges, and though they are effective tools for a given specific location, it is not possible to put them everywhere. Weather radars can cover large areas, but radar readings almost never match the local rain gauge estimates. The U.S. National Weather Service recently upgraded to polarimetric radars, which provide additional measurement types and higher quality data than conventional Doppler radars. Consequently, models that incorporate the new polarimetric data have potential to improve rainfall predictions. Our work attempts to improve existing rainfall prediction techniques by using machine learning tools to model the relationship between rainfall and polarimetric radar observations. We train models on the dataset provided in [1].

1.1 Dataset Description

Polarimetric radars transmit and receive both horizontal and vertical polarizations of electromagnetic pulses alternatively. After each transmitted pulse there is a short listening period, during which the radar receives and interprets reflected signals from the cloud [2]. In addition to measuring the reflected power, polarimetric radars also contrast the horizontal and vertical power returns in different ways (ratios, correlations). These measurements provide information on the size, shape, and ice density of clouds. Some of the fundamental variables measured by polarimetric radars are:

- Reflectivity (Z in dBZ): Reflected power estimate. It is a good indicator of volume density of drops and consequently helps in predicting rain or snow intensity [3].
- Composite Reflectivity (Z_C in dBZ): Gives the highest dBZ (strongest reflected energy) at all elevation scans, not just the reflected energy at a single elevation scan [4].
- Differential Reflectivity (Z_{DR} in dBZ): Ratio of measured horizontal and vertical power returns. It is a good indicator of the drop shape and in turn of average drop size [2].
- Correlation Coefficient (ρ_{HV} unitless): Correlation between reflected horizontal and vertical power returns. It is a good indicator of regions where there is mixture of precipitation types, e.g. rain and snow [2].

[1] This competition was sponsored by the Artificial Intelligence Committee of the American Meteorological Society and hosted by **Kaggle** at URL: https://www.kaggle.com/c/how-much-did-it-rain-ii.

- Specific Differential Phase (K_{DP} in deg/km): Comparison of the returned phase difference between horizontal and vertical pulses. It is a very good estimator of rain rate [2].

Our data comprises 23 columns containing polarimetric radar data and one column containing the corresponding measured rainfall (Q) from a rain gauge. For each gauge, multiple readings from the nearest radar station are taken over the course of one hour. The first column contains an integer ID: multiple rows having the same ID correspond to the set of readings from the same hour for a gauge. The second column contains the minutes elapsed since the beginning of the hour for each reading, and the third contains the distance between radar and rain gauge (in km). Columns 4-23 provide the absolute measured values of $Z, Z_C, Z_{DR}, \rho_{HV}$ and K_{DP} and some percentiles for each of these values in a 5 km × 5 km region from the polarimetric radar [1]. The last column (24) contains the rain gauge reading in millimeters (Q) for a particular ID after one hour. It is same for all rows with the same ID. In the test data, the rain gauge reading is absent and must be estimated. For a detailed explanation of the dataset, refer to [1].

2 Rain Rate Prediction

For this section, we will refer to all reflectivity values (Z, Z_C, Z_{DR}) on a unitless scale as computed by transforming the reflectivity (in dBZ) as:

$$Z \text{ (unitless)} = 10^{\frac{Z \text{ (in dbZ)}}{10}} \tag{1}$$

It is known that radar reflectivity (Z) is a good estimator of rain rate and the relationship between Z and rain rate R (in mm/hr) at any instant is given by the Marshall-Palmer formula [3]:

$$R = \left(\frac{Z}{200}\right)^{\frac{5}{8}} \tag{2}$$

Apart from the Marshall-Palmer formula, there are several other ways to use the other observed values to predict rainrate. For instance the Green's equilibrium axis ratios as given in [5] are as follows:

$$R = 2.47 \times 10^{-2} Z^{0.692} \tag{3}$$

$$R = \text{sgn}(K_{DP}) 46.1 |K_{DP}|^{0.873} \tag{4}$$

$$R = \text{sgn}(K_{DP}) 88.3 |K_{DP}|^{0.982} Z_{DR}^{-1.89} \tag{5}$$

$$R = 7.86 \times 10^{-3} Z^{0.967} Z_{DR}^{-4.98} \tag{6}$$

Their improved versions found through empirical estimation in [5] are as follows:

$$R = 2.62 \times 10^{-2} Z^{0.687} \tag{7}$$

$$R = \text{sgn}(K_{DP}) 54.3 |K_{DP}|^{0.806} \tag{8}$$

$$R = \text{sgn}(K_{DP}) 136 |K_{DP}|^{0.968} Z_{DR}^{-2.86} \tag{9}$$

$$R = 7.46 \times 10^{-3} Z^{0.945} Z_{DR}^{-4.76} \tag{10}$$

Having the rain rate (R in mm/hr) at any point of time during the hour, the total rainfall (Q in mm) during the whole hour can be found by integrating the rain rate over the whole hour. Since we do not have the rain rate during the whole hour, but instead only at several irregularly sampled points for each hour, we will do a time-weighted summation to approximate the integral:

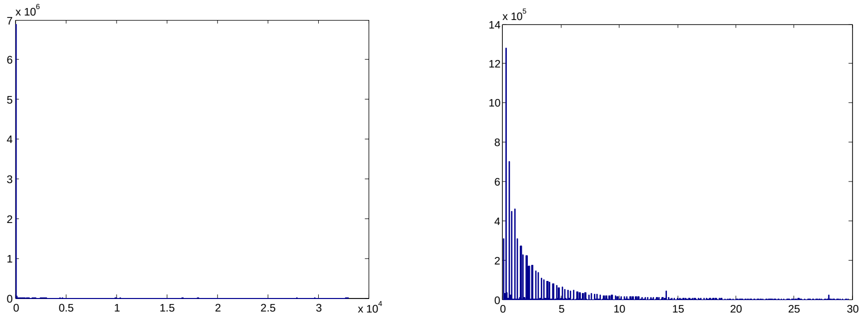
$$Q = \int_t R(t) dt \approx \sum_t R(t) \Delta t \tag{11}$$

3 Data Exploration and Preprocessing

Our dataset contains many missing radar readings, many infeasible rain gauge values and many unlikely to occur radar readings. To explore the nature of data and visualize the effect of various features on the rainfall values, we systematically explored our dataset.

3.1 Missing values and Outliers

The original dataset has 38.42% rows in the training set and 39.46% rows in the test set with all radar readings missing. Several other rows have majority of radar readings missing. Such data points add to the processing load and must be either imputed or filtered out before training. The rain gauge readings contain many outliers and implausible values, e.g. because of gauges getting clogged [1]. This is shown in the histogram of rain gauge readings in training data before filtering (Figure 1a). Such readings negatively impact the training algorithms and need to be removed.



(a) Histogram of rain gauge values pre-filtering (b) Histogram of rain gauge values post-filtering

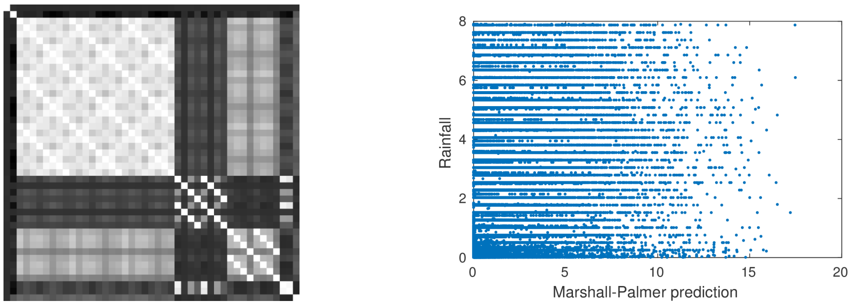
3.2 Filtering and Imputation

Filtering of data was done in the following manner sequentially:

1. All IDs in training set with more than 20% observations missing (NaN) were removed.
2. Q_{95} = 95 percentile of gauge readings (Q) in training set was computed and all IDs in training set with $Q > Q_{95}$ removed.
3. Columns (4-23) of radar readings in both training and test data were checked for missing values. If all values in a row (i.e. at a particular instant) were missing for some ID, then it was assumed that there was no rain at that point in time. If at least one value in that row was available, then any missing value in that row was replaced by the average of other values for that gauge-hour. If all values were missing for column j in ID i, then they were all imputed by full average of column i across the training data.

The thresholds of 20%, 95 percentile and imputed parameter values in the above filtering procedure were found using informal grid searches in conjunction with learning algorithms described in the next section. Figure 1b shows the rain gauge histogram after applying the above filtering steps.

3.3 Exploration



(a) Correlation between processed features (b) True rainfall by Marshall-Palmer

Figure 2: Data exploration visualizations.

We explored the variation of measured rain gauge readings with those predicted by summing-up Marshall-Palmer rain rates computed from reflectivity (Z) values using equations (1)-(2). Figure 2b shows a scatterplot of the Marshall-Palmer model prediction against the true rainfall and illustrates the formidable difficulty of the problem. There is almost no correlation, linear or otherwise, between the Marshall-Palmer prediction and the actual rainfall value. In fact the data is terribly noisy and fairly uniformly distributed for lower values of rainfall. It is also apparent that the gauge reading data are clustered at evenly spaced multiples of 0.254 millimeters or 0.01 inches (presumably the least count of the rain gauges).

3.4 Summarization

The dataset contains samples of radar readings with varying number of time points for each ID (hour). Since few learning algorithms can take data in a variable length format, we need to summarize readings for each hour (ID) into one single fixed-length row. In the following steps, we did not use data from columns 12-15 (ρ_{HV}) and rejected it altogether for all further training experiments because of the lack of summarization techniques involving ρ_{HV} values.

After filtering, all columns containing Z, Z_C, Z_{DR} values (in dBZ) were converted to unitless quantities in both training and test data using equation (1). Then rain rate estimation was performed at each point of time using equations (2)-(10) and new features generated which contained rain rate estimates at the earlier specified time points. We ended up with the new feature matrices having 45 columns.

Finally the data was summarized into one row for each ID (hour), by doing the approximate integral in equation (11) column-wise for every unique ID in both training and test sets. This leads to reduction in the number of rows of the data matrices. The second column ”minutes past” was discarded at this stage. The matrices now contained IDs, distance of radar from rain gauge and various estimates of rainfall during the hour in each column (44 columns) and were used as inputs for training.

3.5 Correlation

At this point we can look at figure 2a which shows the correlation of the resulting columns after summarization. All reflectivity (Z, Z_C, Z_{DR}) columns share a strong positive correlation with each other. In comparison K_{DP} columns are much less correlated internally. Unfortunately, none of the columns display a very strong correlation with the last column containing the rain gauge values and again shows the difficulty of this prediction task.

4 Learning Algorithms

4.1 Marshall-Palmer

Marshall-Palmer rainfall estimates, as computed from Z using equations (1)-(2), provide a good benchmark. The mean absolute error (MAE) on test data using Marshall-Palmer estimates is 24.069. A good training algorithm should be able to improve this error significantly using information in rain rate computed from Z_C, Z_{DR} and K_{DP} using equations (2)-(10).

4.2 Ridge Regression

Our first attempt was to use only Marshall-Palmer estimates of rainfall from Z, Z_C columns in the dataset after preprocessing. Linear regression can combine the 5×5 neighborhood percentiles and column maxima to produce a more accurate estimate. Since these columns show a high correlation to each other, ridge regression was employed to avoid inverting ill-conditioned matrices. The following hyperparameters were selected using grid search with 4-fold cross validation:

- $\{\times 0.75, \times 1, \text{ and } \times 1.5\}$ variants of each constant in Marshall-Palmer equation
- Rain gauge value (Q) outlier cutoff at $\{60, 70, 80, \text{ and } 90\}$ percentile.
- Regularization constant $(\lambda) = \{.0001, .001, .01\}$.

Table 1: Parameters of feedforward neural network

NumInputs	43	λ_{reg}	0.9
NumOutputs	1	Max #Iterations	200
NumHiddenLayers	4	Min Gradient	10^{-10}
Neurons	[500,100,100,10]	CV set size	5%

For each hyperparameter set, models were trained on each $\frac{1}{4}$ of the training data and evaluated on the remaining $\frac{3}{4}$. This approach was chosen to cut down on training time for the 972×4 models. This experiment yielded only about 0.08 difference in cross-validation MAE between the worst and best-performing models, with a mean MAE of 23.945. This showed that results are not sensitive to outlier rejection or Marshall-Palmer parameters, so experiments with later learning algorithms could proceed without conducting a full grid search.

4.3 Feedforward Neural Networks

Since linear regression only combines rainfall estimates in weighted linear sums, it cannot easily approximate the potentially heavily non-linear rainfall model, so we next employed a feedforward neural network for approximating the rainfall predictor more accurately. Neural networks are known to be able to approximate any discontinuous function to arbitrary accuracy given three hidden layers and sufficient number of hidden units per layer [6].

We initially employed a three layer neural network from MATLAB Neural Network Toolbox with 400, 100 and 5 neurons respectively in hidden layers, but realized that more hidden neurons were required to approximate the function well. Adding more neurons leads to huge increase in training time, so we also added an extra hidden layer in addition to increasing the number of neurons slightly in each layer. The connectivity parameters for the final feedforward network were chosen with an informal grid search with repeated training and cross-validation evaluation, because a full grid search was too prohibitive given the huge number of parameters to explore and the limited time.

The network had the following structure:

- **Activation function**: Elliot-sigmoid function was used for hidden layer activations, since it approximates sigmoid without computing exponential, hence speeds up training and evaluation. Output layer had a linear activation for prediction.
- **Performance metric**: We minimize mean squared error (MSE) performance function, since it is differentiable as opposed to mean absolute error (MAE) which is our actual performance metric.
- **Cross validation**: 5% data was held out for cross-validation and the training stopped prematurely if error on CV set increased more than 6 times.
- **Training algorithm**: Gradient descent backpropagation is usually slow in propagating errors due to vanishing gradients and Levenberg-Marquardt has large memory requirements although it offers a nearly quadratic convergence rate. So after initially using LM, we switched to scaled conjugate gradient as a good trade-off between convergence rates and memory requirements.

The final selected parameters are summarized in table 1 and a visualization of the neural network is shown in figure 3.

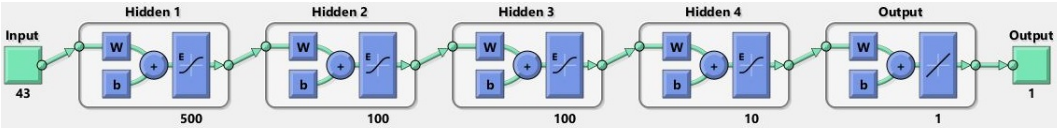


Figure 3: Visualization of feedforward neural network with 43 inputs, 1 output and 4 hidden layers

We finally achieved a MAE of 1.3185 on the training data and 23.897 on the test set with this neural network, which is also the best that we could achieve amongst all the different training and preprocessing algorithms that we tried.

4.4 Outlier prediction

Initial experiments revealed that MAE score of models is dominated by outliers. Models trained on filtered data with lower 85 percentile rain gauge readings, easily achieve scores of 1-2 mm MAE. However, the same models produce much higher MAE scores of about 23.5-24.5 mm on full data.

So we also tried to train models which try to explicitly identify outliers. This turned out to be a daunting task; the data contains many different physically implausible outlier values, including 737 unique values of over 1 meter corresponding to all kinds of radar readings. Given the dramatic effect of outliers on the final model score, we attempted to build a model that could predict outliers successfully using classification algorithms.

We trained several logistic regression models on the full dataset with all outliers. These included a multiclass regression breaking the data into 5-percentile bins and several single-class regression models with various threshold values for the outlier class. Unfortunately, none of these models achieved good accuracy at predicting outliers.

5 Discussion and Conclusion

In this work, we managed to slightly improve the existing Marshall-Palmer rainfall prediction model using a feedforward neural network which also used summarized rainfall estimates from Z_C, Z_{DR} and K_{DP} instead of only using Z. But even after more than 48 hours of training with numerous random restarts and cross-validation terminations, the neural network failed to generalize any better. Given the training time, it is safe to conclude that our model must have been close to the global optimal achievable with our data.

Then the reason for only a small improvement can be attributed to the summarization technique of using a weighted sum to approximate the integral in equation (11). It can be an interesting avenue of further research to try and combine the rain rate estimates using other means or just feeding the raw rain rate and time values to a recurrent neural network. On the other hand, as we demonstrated during data exploration, the data is extremely noisy and full of outliers, so even while employing a recurrent neural network, it might be still infeasible to decrease the MAE beyond a certain limit which calls for better outlier rejection techniques.

References

[1] V. Lakshmanan, A. Kleeman, J. Boshard, R. Minkowsky, and A. Pasch, "The AMS-AI 2015-2016 Contest: Probabilistic estimate of hourly rainfall from radar," 2015. [Online]. Available: https://www.kaggle.com/c/how-much-did-it-rain-ii/data

[2] T. Schuur, "Polarimetric Radar Research - FAQ - Summary," 2003. [Online]. Available: http://cimms.ou.edu/~schuur/radar.html

[3] Wikipedia, "DBZ (meteorology) — Wikipedia, The Free Encyclopedia," 2015. [Online]. Available: https://en.wikipedia.org/wiki/DBZ_(meteorology)

[4] J. G. Tools, "Nws nexrad." [Online]. Available: http://www.desktopdoppler.com/help/nws-nexrad.htm#rainfall%20rates

[5] E. A. Brandes, G. Zhang, and J. Vivekanandan, "Experiments in Rainfall Estimation with a Polarimetric Radar in a Subtropical Environment," *Journal of Applied Meteorology*, vol. 41, pp. 674–685, Jun. 2002.

[6] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, 1989. [Online]. Available: http://dx.doi.org/10.1007/BF02551274