IEEE ICIA 2015
IEEE International Conference on Information and Automation
in conjunction with IEEE International Conference on Automation and Logistics
August 8 – 10, 2015, The Old Town of Lijiang, Yunnan, China
IEEE ICAL 2015

Comparative Analysis of Statistical Models in Rainfall Prediction

Jinghao Niu and Wei Zhang

School of Control Science and Engineering

Shandong University

Outline

- Background
- Prediction Model and Algorithms
- Criterions and Dataset's Features
- Results
- Discussion
- Conclusions

IEEE ICIA 2015

Background

Rainfall Prediction with Data Mining Method (RPDM)

- The predication system applying information extracted from historical records.
- A method that shows obvious advantage in computing cost compared to conventional methods.

Related Work on RPDM

- Many worthwhile studies have been done applying historical data to make rainfall prediction.
- However, their compared results are mainly got from dataset of one specific location, one city for example.
- An examination from larger range of locations may better estimate the performance of the model.

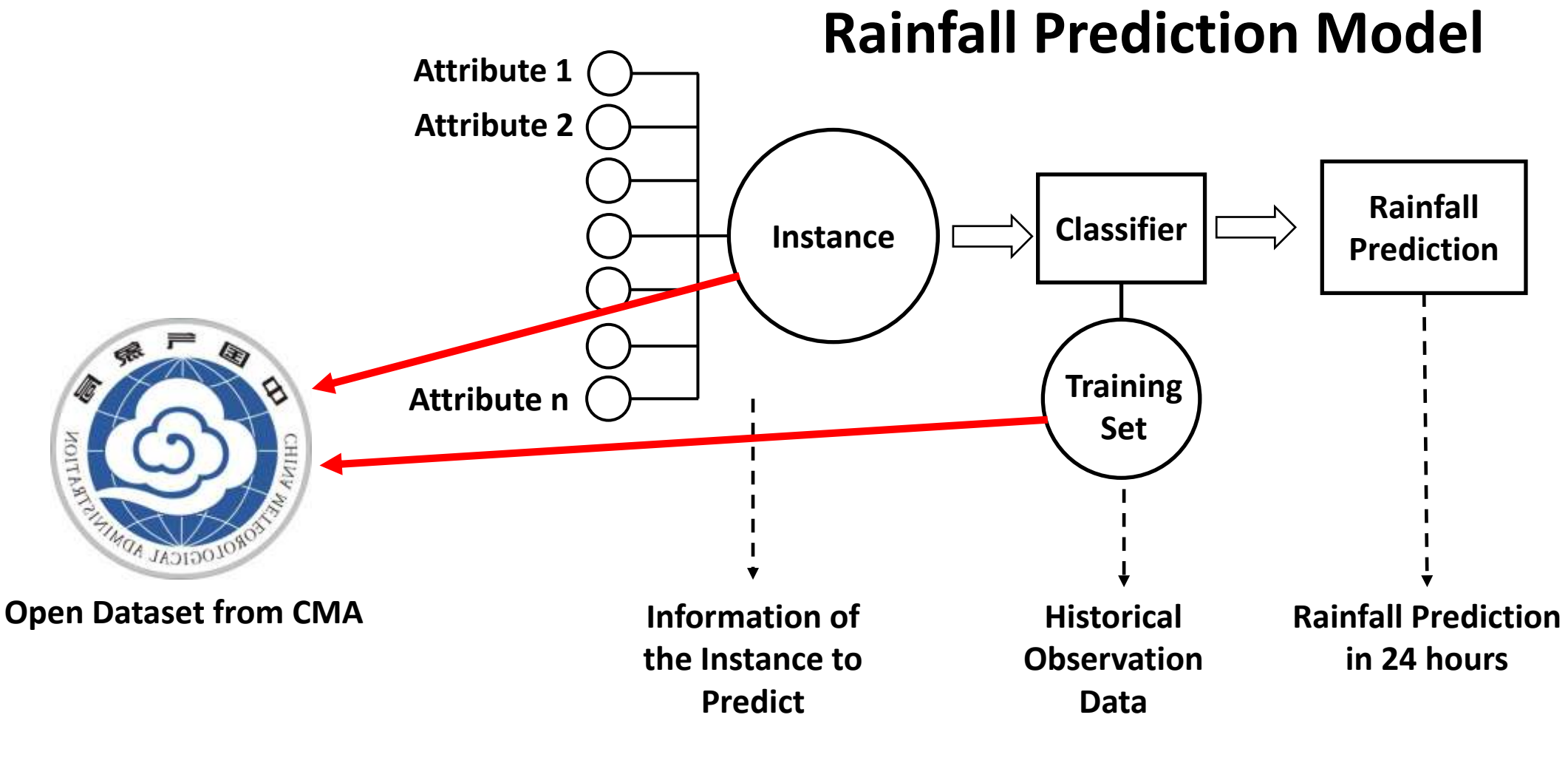
Background

Subsets in This Paper

- 11 representative subsets with obviously different location are chosen from China Meteorological Administration (CMA)'s open dataset.

IEEE ICIA 2015

Prediction Model and Algorithms

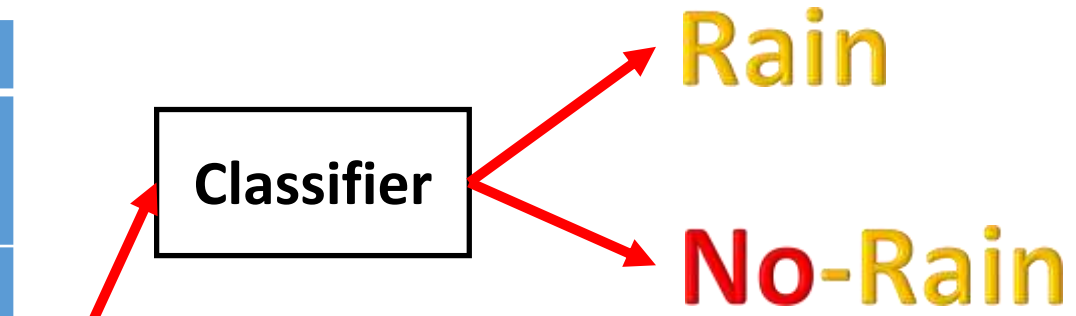


IEEE ICIA 2015

Prediction Model and Algorithms

Input Attributes for the Classifier

Category	Attribute	Remarks
Pressure	The average pressure	0.1hPa
	The highest pressure	
	The lowest pressure	
Temperature	Mean temperature	0.1℃
	Maximum of temperature	
	Minimum of temperature	
Evaporation	Small evaporation	0.1mm
	Large evaporation	
Humidity	Average relative humidity	1%
Wind	Average wind speed	0.1m/s
	Maximum of wind speed	
Sunshine	Total sunshine	0.1hour
Surface Temperature	Average surface temperature	0.1℃
	Maximum of surface temperature	
	Minimum of surface temperature	



Classification Algorithms:

- Naïve Bayes
- Support Vector Machine
- Back Propagation Neural Network

IEEE ICIA 2015

Prediction Model and Algorithms

Details of Classification Algorithms:

- **Naïve Bayes:** one entropy-based discretization method with a stopping criterion based on the Minimum Description Length Principle (MDLP) is applied.
- **Support Vector Machine:**

1. RBF (radial basis function) kernel is assumed for convenient.
2. $\gamma = 2^{-15}$ is the kernel parameter value we found that is suitable for both RO and RR.

- **Back Propagation Neural Network:**

1. Three-layer neural networks with BP algorithm is constructed to make predictions.
2. There are 15 nodes according to attributes from the ground observation.
3. A 16 nodes hidden layer and 2 nodes output layer are constructed to code two classification conditions, rain and no-rain.

IEEE ICIA 2015

Criterions and Dataset's Features

Two Criterions for Evaluating the Prediction Model

- **RO (overall-data-rate):** prediction accuracy upon overall instances in testing set.

$$\mathrm{RO} = \frac{Nc}{Nt} \times 100\%$$

- **RR (rainfall-data-rate):** Only the instance whose real label is rain would be considered and used to calculate the accuracy

$$\mathrm{RR} = \frac{Ncr}{Ntr} \times 100\%$$

- RR is introduced because some classifiers may tend to predict as the value with higher prior probability, which will lead to an ideal RO but terrible RR

IEEE ICIA 2015

Criterions and Dataset's Features

Features of Subsets to Compare

In order to find the relationship between the classification accuracy and some potential different features of observation station, we set comparisons on accuracy with specific feature of stations:

- Latitude (Lat. /N)
- longitude (Log. /E)
- altitude (Alt. /M)
- average temperature (A.T. /℃)
- the prior probability of rainfall (P.P. / %)

Station	Lat.	Lon.	Alt.	A.T.	P.P.
Ha'erbin	45.45	126.46	1423	3.6	50.45
Urumqi	43.47	87.39	9350	5.7	33.67
Beijing	39.48	116.38	313	11.4	41.56
Yinchuan	38.29	106.13	11114	8.5	33.27
Taiyuan	37.47	112.33	7783	9.5	40.56
Xining	36.43	101.45	22952	5.7	56.34
Jinan	36.36	117.03	1703	14.2	40.66
Wuhan	30.37	114.08	231	16.3	42.56
Hangzhou	30.14	120.10	417	16.2	55.34
Guangzhou	23.10	113.20	410	21.8	63.14
Haikou	20.00	110.15	635	23.8	61.14

Results

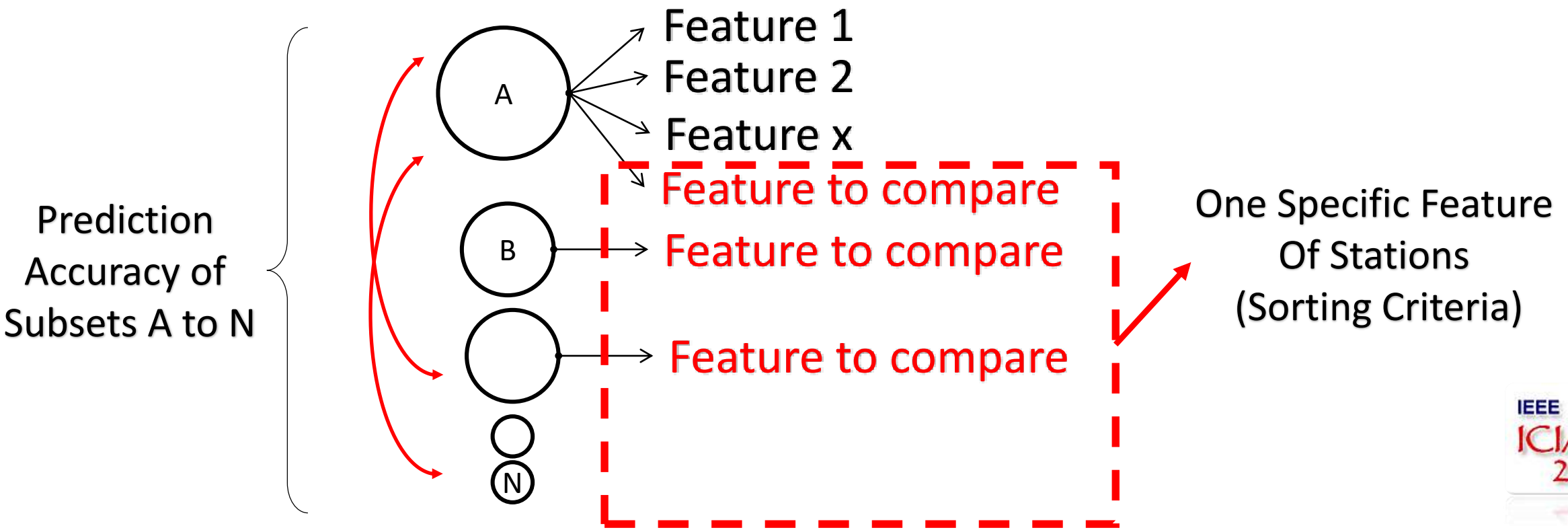
Performance of Prediction Based on RO and RR

Station	NB/%		SVM/%		BPNN/%	
	RO	RR	RO	RR	RO	RR
Ha'erbin	78.04	79.39	79.64	80.53	79.84	80.53
Urumqi	72.06	59.21	80.84	57.89	83.43	69.08
Beijing	73.85	70.16	77.64	66.49	80.04	70.68
Yinchuan	74.45	65.96	75.45	45.74	78.84	61.70
Taiyuan	72.26	66.04	79.04	58.96	78.64	68.40
Xining	71.66	77.30	77.25	83.22	81.24	80.92
Jinan	77.25	70.59	79.64	68.98	82.83	67.38
Wuhan	71.86	87.02	80.64	79.81	76.25	86.54
Hangzhou	76.05	81.95	78.04	84.21	80.84	84.96
Guangzhou	78.04	93.77	82.24	92.13	74.85	90.16
Haikou	68.66	67.22	79.64	87.63	82.24	85.95
Average	74.01	74.41	79.09	73.23	79.91	76.93
Variance	9.29	108.06	3.61	213.96	7.07	93.87

Discussion

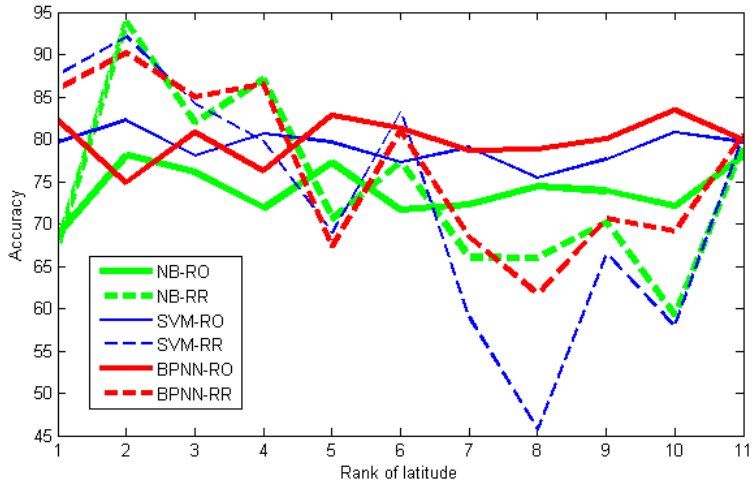
Comparisons of the Accuracy through Stations' Features

In this part, the classification results are reorganized with the sorted feature values, which is for the sake of discovering potential relationships between the classification accuracy and specific potential affecting factor.

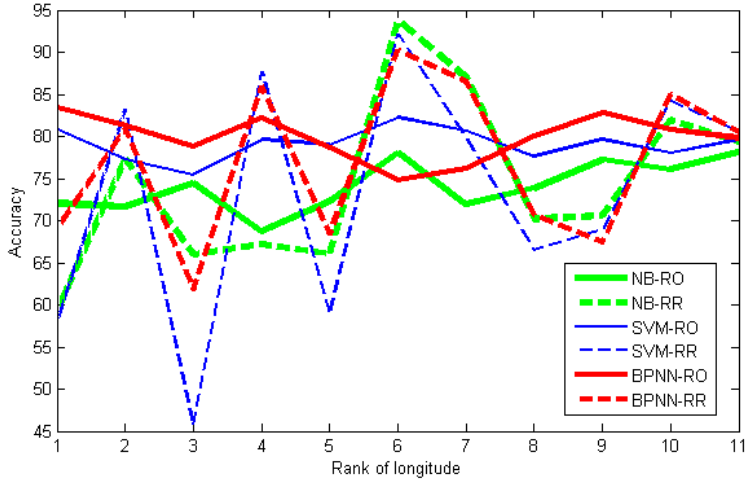


IEEE ICIA 2015

Discussion-1



Sorted by Latitude



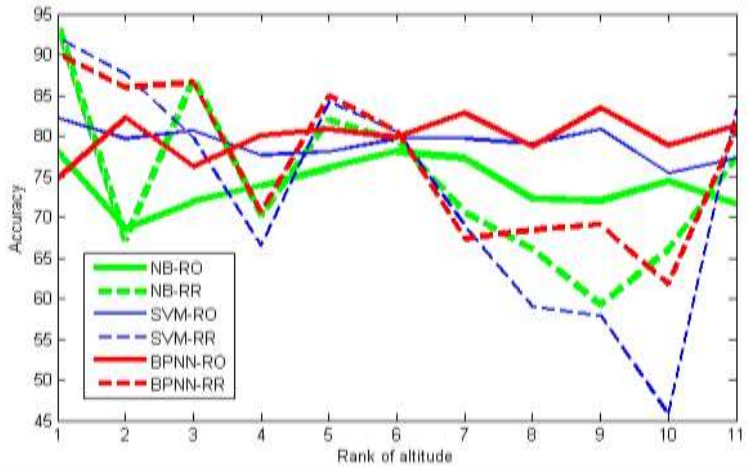
Sorted by Longitude

1- Latitude and Longitude:

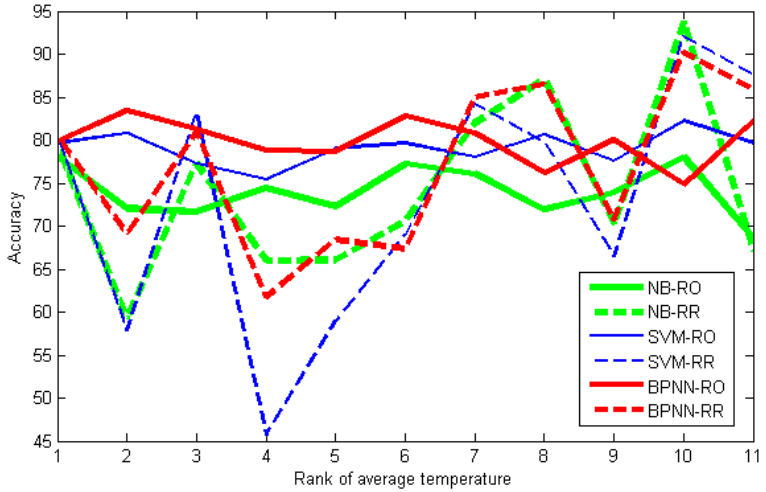
- These two features are directly relative to one observing station's location.
- The fluctuation of RO is visibly smaller than RR.
- For the criterion of RR, relative higher values appear at stations of high latitude locations.
- However, this possible relationship rule does not work for the feature of longitude.

IEEE ICIA 2015

Discussion-2



Sorted by Altitude



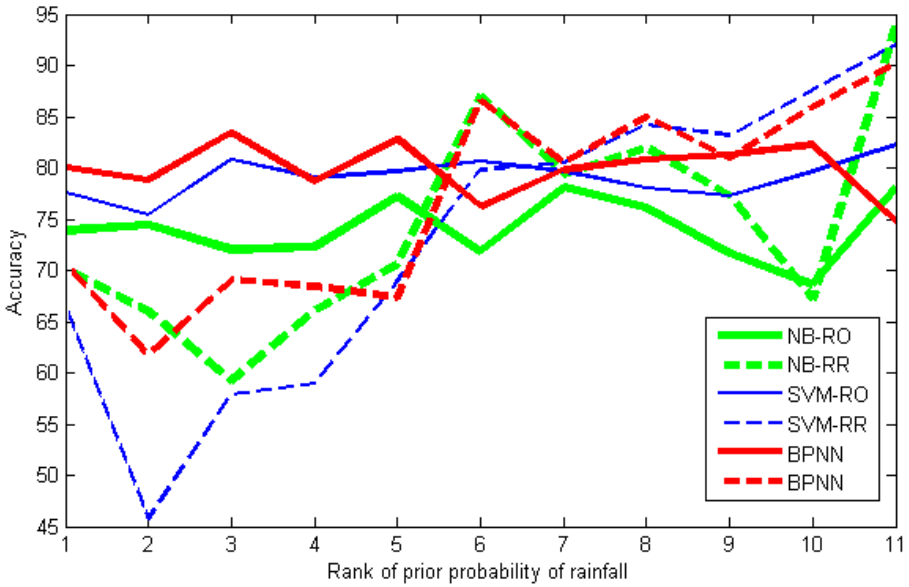
Sorted by Temperature

2- Altitude and Temperature:

- since the error fluctuation from data collection may be more visible compared to the features' affection, RO does not show obvious relationship with these two features as well.
- For RR, we find better performance appears at locations with relative lower altitude and higher average temperature.

IEEE ICIA 2015

Discussion-3



Sorted by The Prior Probability

3- The Prior Probability:

- since the error fluctuation from data collection may be more visible compared to the features' affection, RO does not show obvious relationship with these two features as well.
- For RR, we find better performance appears at locations with relative lower altitude and higher average temperature.

IEEE ICIA 2015

Conclusions

- RR is found to be more sensitive to location change (latitude and longitude) compared to RO, relative higher RR values appear where latitude is relative high.
- In this study, we find better predicting accuracy appear at stations with relative lower altitude and higher average temperature
- As the rise of prior probability, prediction accuracy increases as well. The predicting differences between BPNN and SVM become small at locations with higher prior probability.

IEEE ICIA 2015

Conclusions

- This study gives an analysis on the features of the dataset that affect classification accuracy.
- Since our data is about ground observing records, the relationship we find is just meaningful on rainfall prediction and meteorological data processing.
- However, this study is relative useful as a reference when applying one known location's historical data to predict rainfall.

IEEE ICIA 2015