Proceeding of the 2015 IEEE
International Conference on Information and Automation
Lijiang, China, August 2015

Comparative Analysis of Statistical Models in Rainfall Prediction*

Jinghao Niu and Wei Zhang**

School of Control Science and Engineering
Shandong University
73 Jingshi Road Jinan, Shandong Province, China
niujinghao@outlook.com

Abstract – **Rainfall prediction is an important part of weather prediction. Compared to conventional methods predicting rainfall rate, the approach applying historical records and data mining technology shows obviously advantage in computing cost. Many excellent works have been done attempting to build predicting model with data mining methods, however, most of them just test the predicting accuracy on data set at one specific location. In this paper, we propose two criterions to evaluate the performance of prediction ability. 11 representative subsets with different location are chosen from China Meteorological Administration (CMA)'s open dataset. Every subset is belong to one specific observing station of CMA. Three classification algorithms are tested on our prediction model. We compare varies of combination of observing station feature and classification algorithm (Naïve Bayes, Support Vector Machine and Back Propagation Neural Network). In the end, prediction accuracy of different subsets are sorted through typical features of stations (Latitude, longitude, altitude, average temperature and the prior probability of rainfall) to find their influence on prediction accuracy.**

Index Terms – Rainfall prediction, Data mining, Prediction accuracy, Influence of data set features

I. Introduction

Rainfall prediction with data mining technology, which is different from conventional methods of weather prediction, has received much attention recently. Based on machine learning theory, historical observing data could be exploited to predict the rainfall in future. Compared to other kinds of models, this process of computing is obviously more convenient. Many worthwhile studies have been done applying historical data to make rainfall prediction. For example, Jae-Hyun Seo et al. compared the prediction models' performance using support vector machine (SVM), k-nearest neighbors algorithm (k-NN), and variant k-NN (k-VNN), which generally achieved ideal accuracy on the rain/no-rain in South Korea [1]. Nikam and Meshram built a rainfall prediction model using the data from Indian Meteorological Department, which also worked well with good accuracy [2]. However, on the one hand, since most of attributes extracted from historical data are continuous, the classification accuracy of these present models built with algorithms like Naïve Bayes and C4.5 might be affected distinctly from varies of discretization methods. On the other hand, every group of compared results is mainly got from data set of one specific location, an examination from larger range of locations may better estimate the performance of the model.

The present paper presents a rainfall prediction model applying the ground observation data from China Meteorological Administration (CMA). It sets a comparison among three different algorithms, respectively, Naïve Bayes (NB), Support Vector Machine (SVM) and Back Propagation Neural Network (BPNN). Here the paper proposes two criterions to evaluate the prediction model, the RO (overall-data-rate) and the RR (rainfall-data-rate). 11 representative observation stations are chosen from the data set (194 overall), which are different in aspects of longitude, latitude and the prior probability of rainfall.

This paper is organized as follows: Section II briefly introduces three machine learning algorithms used in this prediction model and the structure of the classifier. In Section III, two criterions to evaluate the prediction model and different features of observations are discussed in detail. Section IV gives the classification results of this model and shows the comparison among different observation stations. Finally in the Section V, the conclusion and some potential improvements for the model are presented.

II. Classification Model and Algorithms

A. Rainfall Prediction Model

In this paper, a rainfall (rain / no-rain) prediction model is built with the technology of data mining. The model extracts information from the ground observation results and exploits it to make rainfall prediction in 24 hours. Compared to conventional methods such as Weather research and forecasting (WRF) model, General Forecasting Model, Seasonal Climate Forecasting Model and Global Data Forecasting Model, this kind of prediction model does not suffer from expensive computing cost [2].

The data set from CMA is consisted of 7 categories of predicting factors and the information of rainfall. 15 attributes are extracted from this 7 categories, which is treated as the input of the classifier. Because there is no rainfall in some specific months, we only choose the data from May to September. And we randomly let 1030 instances in the data set as the training

* This work was supported by the NSFC Grant No. 61203253 and 61233014, Research Found of Outstanding Young Scientist Award of Shandong Province BS2013DX023, Open Program of Jiangsu Key Laboratory of 3D Printing Equipment and Manufacturing 3DL201502, and Program of Key Lab of ICSP MOE China.

** Correspondence should be addressed to Wei Zhang: davidzhangsdu@gmail.com.

978-1-4673-9104-7/15/$31.00 ©2015 IEEE

data, other instances are used for examination. The attributes are showed in detail on TABLE I.

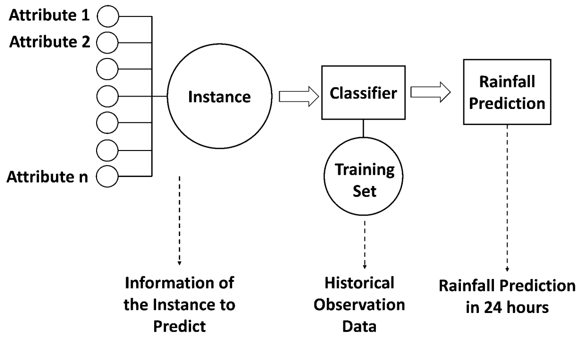
Attribute 1
Attribute 2
Attribute n
Instance
Classifier
Rainfall Prediction
Training Set
Information of the Instance to Predict
Historical Observation Data
Rainfall Prediction in 24 hours

Fig. 1 The structure of rainfall prediction model

TABLE I
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	

B. Algorithms for Classification

For the sake of better estimating to the classification performance, three different algorithms are applied in this prediction model. One of them is Naïve Bayes which needs discretization before classification since attributes form historical observation data set are continuous [3]. The other two algorithms, SVM and BPNN are able to deal with continuous attributes.

1) Naïve Bayes: the attributes used in our classifier model are numeric, we assumed that they follow normal distributions and did a pretest. However, the classification results were terrible, thus the discretization is necessary and could result a higher classification accuracy [4] [5]. We apply one entropy-based discretization method with a stopping criterion based on the Minimum Description Length Principle (MDLP) [6] [7]. For an interval S, E(S) is defined as the entropy of S.

$$\mathrm{E(S)} = -\sum p(Ck,S)log2p(Ck,S) \tag{1}$$

In (1) $p(Ck,S)$ is the proportion of the instances in interval S whose class label is Ck. For one potential splitting point Xj = T that divides S into S1 and S2, the new entropy over S1 and S2 is defined as E(Xj = T).

$$\mathrm{E(Xj = T)} = \frac{|S1|}{|S|}E(S1) + \frac{|S2|}{|S|}E(S2) \tag{2}$$

In (2) |S| is the number of instances in S. The algorithm recursively finds the splitting point to get the smallest new entropy until E(S) – E(Xj = T) < MDLP Gain

$$\text{MDLP Gain}= \frac{\log2(|S|-1)+log2(3^k-2)-[kE(s)-k1E(s1)-k2E(s2)]}{|S|} \tag{3}$$

Where k, k1, k2 is defined as the number of possible class values in specific interval.

2) Support Vector Machine: In this classifier model, RBF (radial basis function) kernel (4) is assumed for convenient [8].

$$\mathrm{K(x,y)} = \exp(-\gamma||x - y||^2) \tag{4}$$

With the evaluating criterions RO and RR, cross-validation is used for the data set to find the best parameter γ. We tested the classifier with different parameters through varies of observation stations [9]. $\gamma = 2^{-15}$ is the kernel parameter value we found that is suitable for both RO and RR.

3) Back Propagation Neural Network: artificial neural network with BP algorithm is tested to be useful in weather forecasting [10] [11]. Three-layer neural networks with BP algorithm is constructed to make predictions, which is a practical method in rainfall prediction [12]. There are 15 nodes according to attributes from the ground observation. In addition, a 16 nodes hidden layer and 2 nodes output layer are constructed to code two classification conditions, rain and no-rain.

III. CRITERIONS AND OBSERVATIONS' FEATURES

A. Two Criterions for Evaluating the Prediction Model

1) RO (overall-data-rate): in this paper, RO is defined as prediction accuracy upon overall instances in testing set. Let the number of testing set be Nt, for every testing instance xj in < x1, x2, … , xNt >, if the output of the classifier, which is either rain or no-rain, is the same as its own rainfall label, this instance would be regarded as one correct prediction. Then, the number of correct prediction through the testing set is Nc.

$$\mathrm{RO} = \frac{Nc}{Nt} \times 100\% \tag{5}$$

2) RR (rainfall-data-rate): Different from RO, RR set one extra limitation to the instance in testing set. Only the instance whose real label is rain would be considered and used to calculate the accuracy. Therefore we get Ncr and Ntr, respectively.

$$\mathrm{RR} = \frac{Ncr}{Ntr} \times 100\% \tag{6}$$

There is a realistic meaning to set RR besides RO. Under some circumstances, the data set being classified may have a strong prior probability to rain (or not), some classifiers may tend to predict as the prior class value, which will lead to an ideal RO but terrible RR. For a more comprehensive evaluate to classifier models, both of RO and RR are taken into consideration.

B. *Features of Observation Stations to Compare*

In order to find the relationship between the classification accuracy and some potential different features of observation station, which means specific meteorology data set in this paper, we set comparisons on accuracy with specific feature of stations. In this paper, Latitude (Lat. /N), longitude (Log. /E), altitude (Alt. /M), average temperature (A.T. /℃) and the prior probability of rainfall (P.P. / %) are these features included. 11 representative observation stations and their features' value are showed in TABLE II.

These features are used to differentiate 11 observation stations in our study.

TABLE II
DIFFERENT FEATURES OF STATIONS

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

IV. CLASSIFICATION RESULTS AND COMPARISONS

A. *General Classification Results of Different Stations*

We choose 11 representative observation stations and test their classification accuracy with NB, SVM and BPNN. For every station, all of 1530 days' historical records, which are from 2005 to 2014, are divided into training set (1030 days) and testing set (500 days). RO and RR are used to represent the performance of every model. TABLE III shows these information in detail.

TABLE III
PERFORMANCE OF CLASSIFIER FOR DIFFERENT STATIONS

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

Basic statistical indicators are computed to evaluate the overall performance of three different algorithms and their sensitivity on change of data set location. BPNN has the best performance both in average RO and RR, compared with NB and SVM. RR is very sensitive to the change of station location, for the RR variance is obviously larger than RO's, no matter what kind of classification algorithm is chosen.

B. *Comparisons of the Accuracy through Stations' Features*

Applying above two criterions, the performance of classifiers could be evaluated, however, factors that affects the performance of prediction accuracy remain to be investigated. Therefore, we choose some subsets that may have different features that affect the classifiers' accuracy. 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.

1) Latitude and Longitude: these two features are directly relative to one observing station's location. However, neither latitude nor longitude alone could show obviously linear relationship with the classification. As we can find in Fig2 and Fig3, 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. The fluctuation of accuracy does not show any trend to follow longitude values.

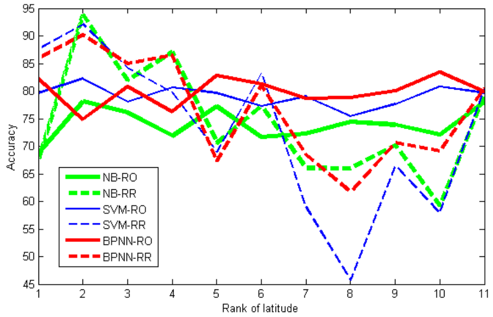

Fig. 2 Accuracy change with feature latitude

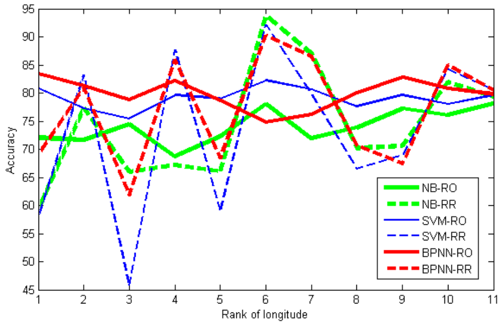

Fig. 3 Accuracy change with feature longitude

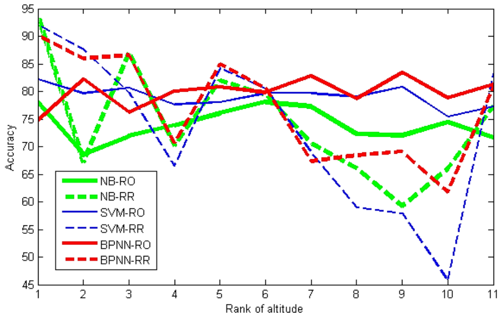

Fig. 4 Accuracy change with feature altitude

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, as Fig.4 and Fig.5 show, we find better performance appears at locations with relative lower altitude and higher average temperature.

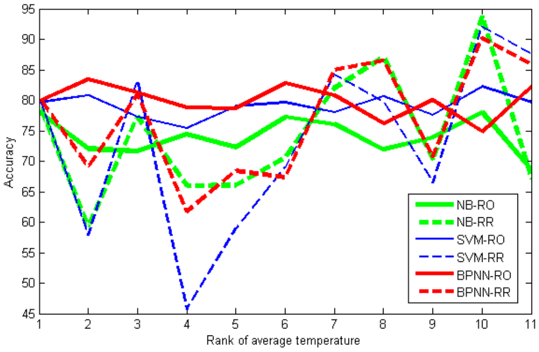

Fig. 5 Accuracy change with feature average temperature

3) The Prior Probability: for the prior probability of rainfall, as it shows in Fig.6, RR is obviously affected by this feature. According to above analysis, feature of latitude is an important factor affecting predicting accuracy. This assertion agrees with our observing on the accuracy change along with the prior probability of rainfall. In China, the city with lower latitude and higher average temperature is more likely with higher prior probability of rainfall, which also leads to better RR in this part of comparison. In addition, at field of high prior probability of rainfall, the performance of SVM-RR and BP-NN show huge similarity. Although the accuracy with NB algorithm is generally worse than SVM and BPNN, factors analyzed above affect its performance more slightly.

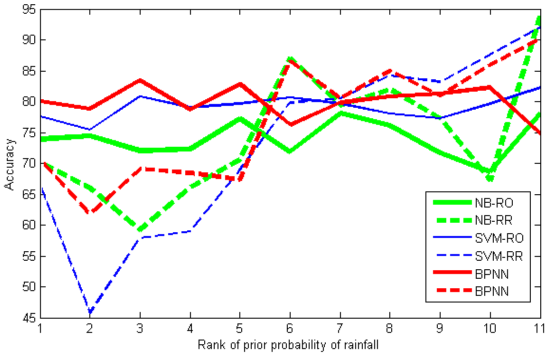

Fig. 6 Accuracy change with the prior probability of rainfall

V. Conclusions

In this paper, we set groups of comparisons of different observing stations and classification algorithms. Applying RO and RR, we evaluate the performance of different rainfall predicting model. In order to find out the relationship between stations' features and predicting accuracy, we draw the relationship figure of classification accuracy through every specific feature.

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. Empirically speaking, higher altitude may lead to a lower average temperature. In this study, we find better predicting accuracy appear at stations with relative lower altitude and higher average temperature. In the end, we test the feature of the prior probability of rainfall, which may most likely be one factor affecting prediction accuracy, since classifiers might tend to predict the class value which has higher prior probability of rainfall. In this part of analyzing, we find 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.

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.

References

[1] J.-H. Seo, Y. H. Lee, and Y.-H. Kim, "Feature selection for very short-term heavy rainfall prediction using evolutionary computation," Advances in Meteorology, vol. 2014, 2014.

[2] V. B. Nikam and B. Meshram, "Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach," in Computational Intelligence, Modelling and Simulation (CIMSim), 2013 Fifth International Conference on, 2013, pp. 132-136.

[3] M. J. Mizianty, L. A. Kurgan, and M. R. Ogiela, "Discretization as the enabling technique for the Naive Bayes and semi-Naive Bayes-based classification," The Knowledge Engineering Review, vol. 25, pp. 421-449, 2010.

[4] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in Machine learning: proceedings of the twelfth international conference, 1995, pp. 194-202.

[5] R. Dash, R. L. Paramguru, and R. Dash, "Comparative analysis of supervised and unsupervised discretization techniques," International Journal of Advances in Science and Technology, vol. 2, pp. 29-37, 2011.

[6] Y. Yang and G. I. Webb, "Discretization for naive-Bayes learning: managing discretization bias and variance," Machine learning, vol. 74, pp. 39-74, 2009.

[7] T.-T. Wong, "A hybrid discretization method for naïve Bayesian classifiers," Pattern Recognition, vol. 45, pp. 2321-2325, 2012.

[8] M. Varewyck and J.-P. Martens, "A practical approach to model selection for support vector machines with a Gaussian kernel," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 41, pp. 330-340, 2011.

[9] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, p. 27, 2011.

[10] D. Sawaitul, K. Wagh, and P. Chatur, "Classification and prediction of future weather by using back propagation algorithm-an approach," International Journal of Emerging Technology and Advanced Engineering, vol. 2, pp. 110-113, 2012.

[11] S. S. Baboo and I. K. Shereef, "An efficient weather forecasting system using artificial neural network," International journal of environmental science and development, vol. 1, pp. 2010-0264, 2010.

[12] J. Long, J. Jian, and Y. Cai, "A short-term climate prediction model based on a modular fuzzy neural network," Advances in atmospheric sciences, vol. 22, pp. 428-435, 2005.