Proceedings of 13th International Conference on Computer and Information Technology (ICCIT 2010)
23-25 December, 2010, Dhaka, Bangladesh

Comparison of Artificially Intelligent Methods in Short Term Rainfall Forecast

Sumi S. Monira, Zaman M. Faisal, H. Hirose
Department of Systems Design and Informatics, Kyushu Institute of Technology, Iizuka, Japan
sumi@ume98.ces.kyutech.ac.jp, zaman@ume98.ces.kyutech.ac.jp, hirose@ces.kyutech.ac.jp

Abstract

Rainfall forecasting has been one of the most scientifically and technologically challenging task in the climate dynamics and climate prediction theory around the world in the last century. This is due to the great effect of forecasting on human activities and also for the significant computational advances that are utilized in this research field. In this paper our main objective is to forecast over a very short-term and specified local area weather using local data which is not always available by forecast center but will be available in the future by social network or some other methods. For this purpose in this paper we have applied three different algorithms belonging to the paradigm of artificial intelligence in short-term forecast of rainfalls (24 hours) using a regional rainfall data of Bihar (India) as a case study. This forecast is about predicting the categorical rainfall (some pre-defined category based on the amount of total daily rainfall) amount for the next day. We have used two classifier ensemble methods and a single classifier model for this purpose. The ensemble methods used in this paper are LogitBoosting (LB), and Random Forest (RF). The single classifier model is a Least Square Support Vector Machine (LS-SVM). We have optimized each of the models on validation sets and then forecast with the optimum model on the out of sample (or test) dataset. We have also verified our forecast results with some of the latest verification tools available. The experimental and verification results suggest that these methods are capable of efficiently forecasting the categorical rainfall amount in short term.

Keywords: Rainfall amount, Artificially intelligent method, Accuracy, Forecast verification.

I. INTRODUCTION

In environmental sciences, forecasting has great socio-economic benefits for society. Forecasts for hurricanes, tornados, Arctic blizzards, floods, tsunami and earthquakes save lives and properties, while forecasts for climate variability such as the El Niño-Southern Oscillation bring economic benefits to farming and fishing and forecast of rainfall has impact on agriculture, reservoir management, power generation/distribution, avalanche forecasting, flood forecasting. Traditionally, rainfall estimates have been mainly derived and forecasted from numerical modeling with both radar and ground observations. With the emergence and evolution of computing, numerical predictions were greatly facilitated by gradually increasing computing capacities; however, although numerical predictions are suitable for long-term forecasts (more than 24 hours) over large areas (several hundred kilometers), short-term mesoscale forecasts in relatively small areas develop the need to find alternatives that give us more specific results.

As an alternative to the numerical forecast, this research presents the methodology from artificial intelligent approaches to short-term rainfall forecasting. This paper studies the possible application of an expert system for rainfall forecasting to short time periods and heavily localized areas. To build the expert system we start analyzing some of the more relevant data-mining ensemble techniques: a) LogitBoosting (LB), b) Random Forest (RF), with a popular single classifier model named, c) Least Square Support Vector Machine (LS-SVM). Our main purpose is making rain predictions in a localized area (using a single meteorological station) and at a very short notice (one day in advance). The novelty of the work lies in the fact that till now to our knowledge no research has been done on rainfall occurrence forecast using all these methods. We selected these algorithms based on their superior performance in classification task. The main objective of the paper is to develop an efficient artificial tool to quick forecast of short term rainfall within a localized spatial region relying on observational data from a single point station with time-series weather records. Although the results obtained by [13] suggest that the use of data obtained from several stations spatially distributed significantly improve the results, this was not the aim of our study.

In our study we start with a brief description of data sources and a discussion of the structure of the models generated with these data, both issues addressed in Section 2. Section 2 also shows the details of the preprocessing of meteorological data that are to be used as inputs in the models. Section 3 describes the rainfall experiments with the available data and Section 4 displays the results; in this section we compare the results of the different models. The last section shows a discussion about the conclusions of this paper.

II. DESCRIPTION OF DATASET AND METHODS OF THE STUDY

In this section we have described about the climate dataset we have used for the rainfall forecast and preprocessing and missing value imputation of the dataset. Followed by a short description of the ensemble methods we have used in this paper.

A. Dataset

In this paper we have used a dataset containing meteorological records averaged over several weather stations

of Bihar region in India. The data was collected from Indian Statistical Institute (Calcutta), which contains rainfall update from 1990 to 2004. The available data provides us measures for a set of parameters for a given day:, such as: evaporation (in mm) denoted as EVA, maximum temperature (in ° C), denoted as MAXT, minimum temperature (in ° C), denoted as MINT, humidity at 8:30 am (%), denoted as HUM830 and humidity at 4:30 pm (%), denoted as HUM430. The output parameter is a numerical variable which gives us the amount of rainfall (in mm) in a day, but we do the next transformation to have a categorical output based on the accumulated rainfall in a day; the categorization is done as follows:

Amount of rainfall	Category
(0 - 1) mm	1
(2 - 9) mm	2
< 10 mm	3

The first category can also be defined as “no rain” category and the last category can be defined as “high rainfall” category. In this paper we have converted forecast of these three categories into two binary forecast problems in the forecast verification. The first binary forecast is “rainfall” vs “no rain” event. Category 2 and 3 together are defined as “rainfall” event and category 1 is defined as “no rain” event. We define this binary forecast problem as RAINFALL. Similarly the other binary forecast problem is “high rainfall” vs “normal rain” event. This is done by defining category 1 and 2 together as “normal rain” event and category 3 is defined as “high rainfall” event. We define this forecast problem as FLOOD. This is defined as such because with such a high amount of rainfall the possibility of sudden flooding is high. As the probability of such high amount of rainfall is rear, this forecast problem is about forecasting a rear event. In addition to these the dataset is scaled and normalized before use. We also log transformed all the predictors. We have used linear scaling computed using the training set, to scale the time series. The scaling step is essential to get the time series in a suitable range, between -1, and 1.

B. CLASSIFICATION MODELS

B1. LogitBoosting (LB):

This ensemble method is from Boosting family. Boosting inherently relies on a gradient descent search for optimizing the underlying loss function to determine both the weights and the learner at each iteration [7]. LogitBoost is a boosting algorithm formulated by Freidman et. al [6] . The LogitBoost fits an additive logistic regression model by the stagewise optimization of the binomial log-likelihood. The LogitBoost framework used in this paper is taken from Dettling and Buhlman [4]. They have restricted the algorithm for decision stumps, which are decision trees with 2-nerminal nodes.

The LogitBoost algorithm for 2-class with class labels:

1) Initialize F(x) =0, wi=1/N, probability estimates p(xi) = ½.
2) Repeat for *m=1,2, ... M*
   a) Compute the working response and weights

$$z_i^m = \frac{y_i^* - p^m(x_i)}{p^m(x_i)(1-p^m(x_i))}$$

$$w_i^m = p^m(x_i)(1-p^m(x_i))$$

   b) Fit the a regression stump fm by weighted least square of zi to xi using the weights w_i.

$$f^m = \text{argmin}_m \sum_{i=1}^{N} w_i^m (z_i^m - f(x_i))^2$$

   c) Updating the weights and the classifier output values

$$F^m = F^{m=1} + \tfrac{1}{2} f^m$$

$$p^m = \frac{1}{1+e^{-2yF^m}}$$

$$C^m = sign(F^m)$$

3) Combine the outputs Cm by weighted majority voting.
4) Where $y^* = \frac{y+1}{2}$, so that $y^* \in \{0,1\}$.

In this algorithm the loss function is the binomial log-likelihood. It increases linearly for strongly negative margins [8]; for this it is more robust in noisy problems.

B.2 Random Forest (RF):

Breiman [1] proposed random forests, which add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers and is robust against overfitting [1]. In addition, it is very user-friendly in the sense that it has only two parameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values. Figure 1 shows how the training data is sampled to create an in-bag portion to construct the tree, and a smaller out-of-bag portion to test the completed tree to asses its performance. This performance measure is known as the out-of-bag error estimate.

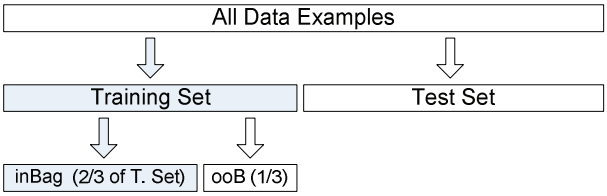

Fig. 1: Breakdown of Data used to build trees

Construction of a Tree:

1. Randomly sample with replacement (bootstrap) the training set and select 2/3 of data to be used for tree construction (inBag).
2. Choose a random number of attributes from the inBag data and select the one with the most information gain to comprise each node.
3. Continue to work down the tree until no more nodes can be created due to information loss.
4. Compute out-of-bag error estimates by running (ooB) dataset through tree and measuring its correctness.

B.3 Least Square Support Vector Machine (LS-SVM):

A modified version of SVM classifiers, Least Squares SVMs (LS-SVMs) classifiers, was proposed by Suykens and Vandewalle in [21]. A two-norm was taken with equality instead of inequality constraints so as to obtain a linear set of equations instead of a QP (Quadratic Programming) problem in the dual space.

Least Squares SVM (LS-SVM) applies the linear model, where ω the weight vector, and b is the bias of the linear model. ω and b are estimated by minimizing the following objective function:

$$L(\omega, b) = \sum_{i=1}^{n} \| (f(x_i) - y_i) \|_2^2 + C \| \omega \|_2^2$$

where y is the vector of class labels, and $C > 0$ is the regularization parameter.

III. EXPERIMENT SETUP

In this section we have described the setup of the experiment of this study. Then we have given a precise description of the metrics used for forecast verification in this paper.

C.1 DESCRIPTION OF EXPERIMENT SETUP:

The classification models we have used in this paper are sophisticated and can perform variably without the proper selection of parameters. For example the LB can perform better with larger number of iterations than with smaller number of iterations. So it is very important that before applying the models in out of sample forecast they are trained with optimized parameters. To select their optimum parameters for the forecast problem we have designed a grid with their parameters; for LB the parameters are 1) number of iterations or number of trees. For RF the parameters we optimized are number of trees and number of features randomly selected. For LS-SVM we have opted for the radial basis function kernel (RBF) and the other parameter is the kernel width or sigma (σ). In Table I we have given the values of the parameters of each method and total number of points in the grid.

In this paper we have conducted our experiment in three phases, a) Training phase, b) Validation phase and c) Out of sample (Test) phase. For this purpose we have partitioned our data in two equal parts, training set and test set. The training-validation period is from 1990~2000 and the test phase is from 2001~2004. The training and validation is repeated 20 times using moving block cross-validation.

TABLE I: Parameters values of the ensembles for optimization.

Methods	Parameter values		Total grid points
LB	# of trees: 50, 100, 150, 200 and 200		5
RF	# random features 2,3,4,5	# of trees: 50, 100, 150, 200 and 200	4x5=20
LS-SVM	Sigma: 0.01, 0.1, 1, 2, 4, 8		5

In each block a fixed percentage of observations (67%) in each repetition are carefully selected so that the training and validation sets are as much as independent. This is done so that the data records adjacent to or near the omitted observation(s) will tend to be more dissimilar to them than randomly selected ones, so the omitted observation(s) will not be more easily predicted than the uncorrelated future observations they are meant to simulate. The final phase i.e., test phase is conducted with the unseen data (test data) and the ensemble method with the optimized parameters.

C.2 METRICS FOR VALIDATION PHASE AND FORECAST VERIFICATION:

In the validation phase we have emphasized on selecting the parameters which enable the models to have a better accuracy and also having fair disagreement in the prediction of the resamples. The metrics are Fraction Correct (FC)/ Accuracy and by Cohen's Kappa [2] coefficient (k).

As we are doing binary forecast the metrics for forecast verification is a little different than the usual evaluation metrics. To do that we have utilized some useful statistics available from signal detection theory which are now frequently used in the climatology. Signal detection theory was first applied to the verification of me-

TABLE II: Schematic contingency table for categorical forecasts of a binary event. The symbols a–d represents the different number of events observed to occur in each category.

Event Forecast	Event Observed		
	Yes	No	Total
Yes	a (hits)	b(false alarm)	a+b
No	c (miss)	d(correct rejection)	c+d
Total	a+c	b+d	a+b+c+d =N

teorological forecasts in the pioneering studies of Mason ([14], [15]) and provides a universal framework for evaluating the joint probability distribution of forecasts and observations. In Table II we have represented a contingency table for the binary event forecast. The numbers in each category will be represented by the symbols given in Table II. In this study, the columns

TABLE III: Description of the Metrics used in this paper.

	Applied in Phase	Equation for computing	Range	Comments
FC	Validation, Test	a+d/N	0 to 1 perfect = 1	Simplest measure, but very sensitive to use in rear climatic events.
k	Validation	NA	-1 to 1 fair = 0.21~0.40 [11]	Measure the amount of agreement among the predictions over the resamples.
H	Test	a/(a+c)	0 to 1 perfect = 1	Very sensitive to the climatological frequency of the event.
F	Test	b/(b+d)	0 to 1 perfect = 0	same.
FAR	Test	b/(a+b)	0 to 1 perfect = 0	same.
Bias	Test	(a+b)/(a+b)	0 to inf. perfect = 1	Indicates whether the forecast system has a tendency to underforecast (BIAS<1) or overforecast (BIAS>1) events.
OR	Test	ad/bc	0 to inf. perfect = inf.	Measures the ratio of the odds of making a hit to the odds of making a false alarm.
ORSS	Test	(ad+bc)/ (ad-bc)	-1 to +1 perfect = 1	It is a, ‘‘measure of association’’. Independent of the marginal totals.
PSS	Test	(ad-bc)/(a+c)(b+d)	-1 to +1 perfect = 1	Uses all elements in contingency table. Does not depend on climatological event frequency.
EDS	Test	(logF-logH)/(logF+logH)	-1 to +1 perfect = 1	Measures the association between forecast and observed rare events.
SEDS	Test	[{log((a+b)/N)+ log((a+c)/N))}/{log (a/N)}]-1	-1 to +1 perfect = 1	It is equitable and symmetric in nature than EDS.
AUC	Test	NA	0 to 1 perfect = 1	For binary prediction problem, this measure computes the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

are used to denote the observed variable, while the rows are reserved for the predicted variable. We shall define the metrics (statistics) based on the symbols of Table II. The computed metrics are fraction correct (FC), hit rate (H), false rate (F) and false alarm ratio (FAR), odds ratio (OR) [19] and degrees of freedom for calculating OR (n_h). We have also computed some skill scores to check the skill of the methods in out of sample forecast; the scores we computed are, Bias, Pierce Skill Score (PSS), Extreme Dependency Score (EDS) [20], Odds Ratio Skill Score (ODSS) [19], Symmetric EDS (SEDS) [9] and Area under the ROC Curve (AUC) [5]. In Table III we have stated how these metrics are calculated and their limit.

C.3 SOFTWARE:

We have used the free statistical software R version 2.10-1 [18] to compute all the results of this paper. We have executed LB using the packages caTools [22] and

ada [3]. The RF is executed using the package random Forest [12]. The grid computation is designed using the package caret [10]. All the forecast verification metrics are computed using the package verification [17]. The cross-validation is designed using the package ipred [16].

IV. RESULTS AND DISCUSSION

We have presented the validation results of the models first, and then the test results are presented. In Table IV the optimum parameters of the classification models obtained in the validation phase are given with the value of the metrics. We see that the LB has the highest accuracy with lowest kappa value, this indicate that there are lesser agreement between the predictions of the LB method over resamples.

TABLE IV: Optimum parameters and evaluation metric values of the models in validation phase

Methods	Optimum Parameter values		Metric Value	
			FC	k
LB	# of trees: 150		**0.8569**	0.2662
RF	# random features: 2	# of trees: 150	0.8183	0.3487
LS-SVM	Sigma: 0.1		0.8162	0.2709

In the out of sample or test phase, the FLOOD forecast problem can be viewed as a rare event forecasting. So some of the metrics we have calculated can be misleading. Mainly because the other measures (FC, H, F and FAR) are heavily influenced by the frequency of the climatological events, so inference on extreme events based on these measures will be rather biased. Stephenson in [19], for verification of rare event forecasting proposed to use Bias, OR, PSS and ORSS. Recently in [20] Stepehnson et.al proposed a measure EDS and Hogan in [9] proposed SEDS, these two measures are till now quite robust for verification of the rare event forecast. In addition to these we have also computed the AUC of models for both the problems. As we know higher AUC is desirable for binary prediction problem and this measure is now used more frequently than accuracy in binary prediction problems. The RAINFALL forecast problem can be viewed as an as usual forecast problem, so for this case values of all the measures should be taken into account. We have presented all the test results of both forecast problems in Table V. The most desirable (based on the range and comments presented in Table III) values are marked bold.

TABLE V: Optimum parameters and evaluation metric values of the models in validation phase

Metrics	FLOOD FORECAST			RAINFALL FORECAST		
	LB	RF	LS-SVM	LB	RF	LS-SVM
FC	0.9162	0.8897	0.8932	0.8594	0.8913	0.8584
H	0.5271	0.5312	0.3543	0.4717	0.3315	0.4461
F	0.0056	0.0282	0.0468	0.0419	0.0458	0.0317
FAR	0.4167	0.2277	0.5287	0.2824	0.3422	0.2500
Bias	0.7429	0.7768	0.7411	1.076	1.0905	1.112
OR	16.182	11.763	10.704	31.639	18.392	20.221
ORSS	0.8836	0.8433	0.8291	0.9387	0.8969	0.9058
PSS	0.2777	0.3193	0.2935	0.4503	0.4036	0.3664
EDS	0.3891	0.3882	0.3568	0.4173	0.3414	0.2826
SEDS	0.388	0.4651	0.4459	0.6065	0.5355	0.5293
AUC	0.5389	0.6596	0.6468	0.7252	0.7018	0.6832

From Table V we see that the though the accuracy is very high for LB in FLOOD forecast problem but it has a poor AUC value, on the contrary though RF has lower accuracy than LB but it has achieved better AUC than both LB and LS-SVM. Regarding the measured suitable to judge the rare event forecast we can say that RF has higher bias value (more closer to 1 than other methods), higher PSS and SEDS value. On the contrary LB has higher OR, ORSS and EDS value. Performance of LS-SVM is also satisfactory for this purpose; its values are very close to the values of both RF and LB. For the RAINFALL forecast problem, the performance of LB is far better than both RF and LS-SVM. It has higher AUC, SEDS, EDS, PSS, ORSS, OR, better bias (more closer to 1 than other methods) and H. So these results suggest that the classification methods from the artificial intelligence paradigm can be successfully utilized to forecast climatic events.
to obtain a rainfall forecast. We emphasized more on predicting correctly the occurrence of the events than non-occurrence of the events. Optimization of the parameters is conducted to get the best performance from the classification models. Also we have checked the verification of the forecast of these methods with verification metrics proposed very recently and suitable for both the forecast problem. For the very high rainfall forecast problem the performance of RF is better than other two methods based on the scores of the metrics. But for the usual rainfall occurrence problem, the LogitBoosting ensemble method produced best forecast results. For the future work these methods can be empir-

V. CONCLUSION

In this paper a novel application of three artificial intelligent methods are studied. The methods are applied to forecast rainfall occurrence and very high rainfall on short term over a localized region. The system we have studied may be useful in some areas; for example, to do forecasting with personal home meteorological stations or in regions in which local rainfall forecasting can be critical and will be available in the future by social network or some other methods. In addition, this type of system can provide us a relatively fast and simple way

ically compared with the usual forecasting methods (Neural Network, ARIMA).

REFERENCES

[1] Breiman, L. Random Forest. Machine Learning. Vol.25, pp. 5-32, 2001.

[2] Cohen, J. A coefficient of agreement for nominal scales, Educational and Psychological Measurement. Vol.20, No.1, pp. 37–46, 1960.

[3] Culp, M., Johnson, K., Michailides, G. ada: An R Package for Stochastic Boosting. Journal of Statistical Software, Vol. 17, No. 2, pp. 1-27, 2007.

[4] Dettling, M., Buhlman, P. Boosting for tumor classification. Bioinformatics. Vol. 19, pp. 1061-1069, 2003.

[5] Fawcett, T. An introduction to ROC analysis. Pattern Recognition Letters, Vol. 27, pp. 61–874, 2006.

[6] Friedman, J., Hastie, T. and Tibshirani, R. Additive logistic regression: a statistical view of boosting. Annals of Statistics, Vol. 28, pp. 337-407(with discussion), 2000.

[7] Friedman, J. Stochastic gradient boosting. Computational Statistics & Data Analysis. Vol. 38. pp. 367-378, 2002.

[8] Hastie, T., Tibshirani, R. and Friedman, J. The Elements o f Statistical Learning. Second Edition, Springer, New York, 2009.

[9] Hogan, R. J., Connorand, J. O., Illingworth, A. I. Verification of cloud-fraction forecasts. Quarterly Journal of The Royal Meteorological Society, vol.135, pp. 1494–1511, 2009.

[10] Kuhn, M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software, Vol. 28, No. 5, pp. 1-26, 2008.

[11] Landis, J. R., Koch, G. G. The measurement of observer agreement for categorical data. Biometrics, Vol. 33, pp. 159–174, 1977.

[12] Liaw, A., Wiener, M. Classification and Regression by randomForest. R News, Vol. 2, No. 3, pp. 18-22, 2002.

[13] Liu, J.N.K., Lee, R.S.T. Rainfall Forecasting from Multiple Point Source Using Neural Networks. In: Proc. IEEE Int'l. Conf. Systems, Man, and Cybernetics (SMC 1999), Vol. II, pp. 429–434, 1999.

[14] Mason, I. B. Decision-theoretic evaluation of probabilistic predictions. Proc. WMO Symp. on Probabilistic and Statistical Methods in Weather Forecasting, Nice, France, WMO, pp. 219-228, 1980.

[15] Mason, I. B. A model for the assessment of weather forecasts. Australian Meteorology Magazine, Vol. 30, pp. 291–303, 1982.

[16] Peters, A., Hothorn, T., Lausen, B. ipred: Improved Predictors. R News, Vol. 2, No. 2, pp. 33-36, 2002.

[17] Pocernich, M. Contributed R package: verification. Version 1.31, 2010.

[18] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2010.

[19] Stephenson, D. B. Use of the "odds ratio" for diagnosing forecast skill. Weather Forecasting, Vol. 15, pp. 221-232, 2000.

[20] Stephenson, D. B., Casati, B., Ferro, C.A.T., Wilson, C. A. The extreme dependency score: a non-vanishing measure for forecasts of rare events. Meteorological Application. Vol. 15, pp. 41-50, 2008.

[21] Suykens, J. A. K., Vandewalle, J. Least squares support vector machine classifiers. Neural Processing Letters, vol. 9, pp. 293–300. 1999.

[22] Tuszynski, J. Contributed R package: caTools, Version 1.10, 2009.