International Journal of Advances in Life Science and Technology
2020 Vol. 4, No. 1, pp. 11-18
ISSN(e): 2313-8335
ISSN(p): 2412-3803
DOI: 10.18488/journal.72.2020.41.11.18

COMPARABLE INVESTIGATION FOR RAINFALL FORECASTING USING DIFFERENT DATA MINING APPROACHES IN SULAYMANIYAH CITY IN IRAQ

Sherko H. Murad[1+]
Yusra Mohammed M. Salih[2]

[1,2]*Computer Science Department, Kurdistan Technical Institute, Sulaymaniyah, Iraq.*
[1]*Email: sherko.murad@kti.edu.krd Tel: 07714155324*
[2]*Email: yusra.mohammad@kti.edu.krd Tel: 07501151447*

(+ Corresponding author)

Article History
Received: 10 February 2020
Revised: 12 March 2020
Accepted: 16 April 2020
Published: 13 May 2020

Keywords
Rainfall forecasting
Data mining approaches
Artificial neural network
Support vector machine
Naïve bayes
Weather prediction.

ABSTRACT

Weather prediction is a critical assumption in weather forecasting. Weather prediction and has been one of the major scientifically and technologically demanding issues worldwide in the last century. The most significant parameter in a hydrological model is Rainfall. The meticulous Rainfall forecasting is one of the major demanding in the atmospheric research. The factors such as pressure, temperature, humidity, wind speed, mean sea-level etc. are used for rainfall forecasting. This study evaluates multiple classifiers such as Artificial Neural Network (ANN), Naïve Bayes and Support Vector Machine for rainfall prediction in Sulaymaniyah city and describes which one is most suitable to predict the precipitation. The dataset has been collected from weather forecast department in Sulaymaniyah city. Pre-processing technique such as cleaning and normalization processes is used for effective prediction. The data mining approaches are evaluated and the Performance is analyzed regarding precision, recall and f-measure with numerous ratios of training and test data.

Contribution/Originality: This paper contributes the first logical analysis for the rainfall forecasting in Sulaymaniyah city. The dataset collection is based on the local forecasting department, Weather Forecast department, in the city. This study tests several supervised learning approaches including (ANN), (SVM), and (NB) to perform a comparative analysis concerning their ability for rainfall prediction in the region.

1. INTRODUCTION

Rainfall is a complicated climatic technique, which is depend on some particular structures such as moisture and wind in a specific location and it is not easy to predict. Because of the noticeable random characteristics of rainfall series, they are often defined by a hypothetical process. Accurate rainfall predicting will aid in evaluating drought and flooding situations in advance. consequently, it is important to have a great model for rainfall predicting [1].

Data mining is a set of methods used to draw out unknown portion of information from the large database source. Analyzing data statistically and extract or derive such rules that can be used for prediction using Data mining approaches [2]. Currently it is being used in many domains such as stock market, sports, banking section, etc. Sheikh, et al. [3] Researchers have now realized that data mining can be used as a mechanism for weather forecasting as well.

There are many related studies done by researchers on several classification approaches for Rainfall Prediction including Neural Network, K-Nearest Neighbor and Naive Bayes, Support Vector Machine and others. Locations such as Korea, Malaysia, South Africa and others are investigated for rainfall forecasting using Several proposed data mining approaches. Hence, there is a necessity to investigate new locations based on the data mining approaches in order to classify the best performance in terms of rainfall forecasting in a specific location. For this reason, we explore a new location such as Kurdistan region of Iraq for rainfall prediction and especially Sulaymaniyah city.

This study tests several supervised learning approaches including Artificial Neural Network (ANN), Support Vector Machine (SVM), and Naïve Bayes (NB) to perform a comparative analysis concerning their ability for rainfall prediction in Sulaymaniyah city in Kurdistan region, Iraq. The data has been gained from the weather forecast department in Sulaymaniyah city. Then the data is analyzed, filtered and normalized so as to efficiently train the data mining approaches to predict the rainfall.

The rest of the paper is organized as follows: Section 2 is the related work whereby it covers some of the research studies that have been previously studied for rainfall forecasting using several data mining techniques based on different dataset. The Material and Approaches are presented in section 3. The Result and Discussions are given in section 4. Lastly, the conclusion and future work are in section 5.

2. LITERATURE REVIEW

In research area, Literature survey plays a very important role which provides the essential knowledge about the study and its background. There are various methodologies used for rainfall prediction using machine learning techniques.

From the correlated works, Supervised machine learning approaches such as Support Vector Machine (SVM), Naïve Bayes (NB) Neural Network (NN) have been commonly used for rainfall forecasting [4]. The idea behind supervised learning approaches is choosing a suitable approach with suitable features.

Researchers in Zainudin, et al. [4] implemented several classifiers such as Support Vector Machine, Decision Tree, Neural Network, Naïve Bayes, and Random Forest for rainfall prediction in Malaysia. The dataset was gained from several stations of Selangor, Malaysia. Dataset is analyzed and compared, according to the outcomes, Random Forest achieved better accuracy on large test data as it appropriately classified large number of instances.

In Aftab, et al. [5] researchers performed five different classifier for rainfall prediction in Lahore city in Pakistan. Three accurateness measures: Precision, Recall and F-measure are used to evaluate the performance of data mining approaches. According to the outcomes, the performance for the no-rain class is better than for rain class because of having missing values and the lack of climate features in the dataset.

In Mishra, et al. [1], Artificial Neural Network (ANN) with Back Propagation algorithm and Levenberg-Marquardt training function was used for developing one-month and two- month afterward forecasting models for rainfall prediction. A monthly rainfall data of 141 years is collected from several weather stations in north India. According to the proposed models, ANN showed optimistic results.

Historical data Collected from the Metrological Department in India and Bayesian algorithm used for rainfall prediction is proposed by Nikam and Meshram [6]. Based on the used features, mean sea level, pressure level, temperature, vapor pressure, wind speed, and relatively humidity, the Bayesian algorithm was trained on the data. The performance of the model is acceptable, but it can be improved by using large dataset.

More recently, several approaches have been proposed for rainfall forecasting for many locations such as Korea, China, South Africa and others [7]; [8]. The current techniques for rainfall prediction including Naive Bayes, Support Vector Machine, Neural Network, and others [9].

3. MATERIAL AND APPROACHES

Similar work has been proposed for rainfall forecasting using different data set and location. This study purposes the performance analysis for rainfall prediction in Sulaymaniyah city using several data mining approaches. Weka, one of the expansively used Data mining software [5], is used for the classification performance in this study. The classification background is illustrated in the Figure 1.

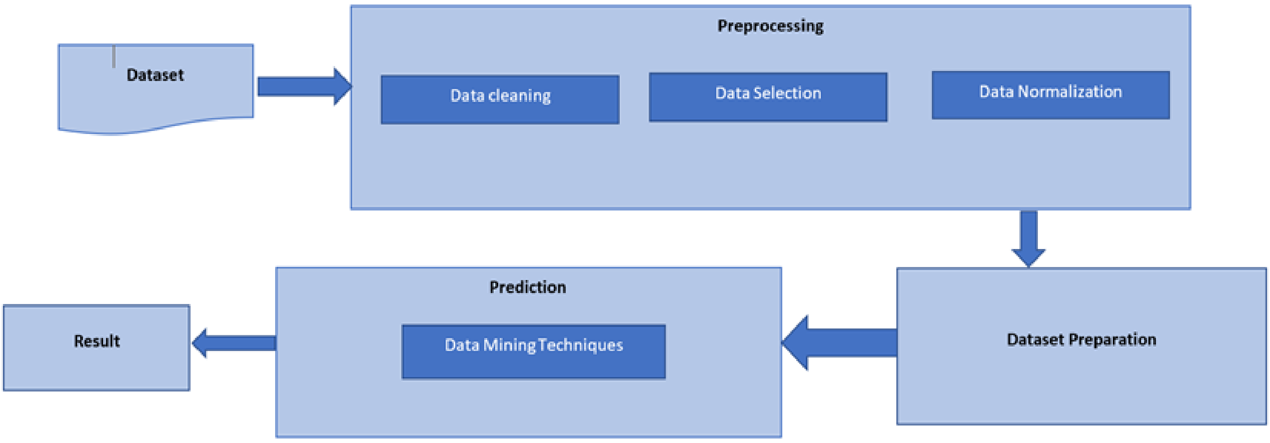

Figure-1. The workflow of the study.

3.1. Data Acquiring and Preprocessing

The data used for this work gathered from the Sulaymaniyah Weather Forecast department in Sulaymaniyah city. The case data covers the period of 2013 to 2018. Data cleaning, and Data Selection, were adopted in this stage.

In Data cleaning phase a reliable data model format was established in a way which missing data are examined, identify the duplicated data, and tidying out of inadequate data. Essentially, the cleaned data were altered into a suitable format for the data mining approaches.

In the Data Selection stage, the Climatological dataset has ten attributes and using the last attribute as a predictor. The intensity of the rainfall is represented by ten different classes, based on the maximum and minimum temperature of the day, relative humidity and rainfall. The attribute sample for rainfall is generated as if the rainfall amount is greater than 0, The attribute sample for Not rainfall is generated as if the rainfall amount is equal to 0 [10]. Measurements for the "Rain" and "No Rain" are depicted in Table 1:

Table-1. Descriptive rainfall measurement terms.

Descriptive Term used	Rainfall amount (mm)
No Rain	0.0
Rain	> 0.0

Source: Narvekar and Fargose [10].

3.2. Dataset Preparation

The collected dataset consists of 2157 instances; all data should be normalized to prepare for a suitable data for the proposed methods. The average of maximum and minimum temperature is selected to have one attribute temperature. Table 2 shows the attributes and target value for the dataset.

Table-2. The distribution of predictor and target variables.

Target Variable	Selected Variables
Precipitation (Rainfall)	Speed Wind
	Direction Wind
	Humidity
	Temperature
	Vapor
	Sunshine
	Cloud Cover
	Sea Pressure
	Station Pressure

After preprocessing stage of the dataset, the collected variables are divided into two groups "Rain" assigned as "Yes" and "No Rain" assigned as "No".

The selected data samples are transferred to a spreadsheet file for further processing to be suitable for data mining approaches. The data set were normalized to minimize the effect of scaling on the data and saved as a Commas Separated Value (CVS) file format.

3.3. Data Mining Approaches and Proposed Methods

In this stage, Artificial Neural Network (ANN) algorithm with percentage split, Support Vector Machine (SVM) with cross validation, and Decision Tree algorithm were examined to analyze the Climatological datasets. Finally, the result of each algorithm is evaluated and the best one has been chosen to predict the weather as "Rain" or "No Rain".

3.3.1. Artificial Neural Network (ANN)

A neural network is a computational structure inspired by the study of biological neural processing [7]. One of the feed forward ANN models is a Multi-Layer Perceptron model (MLP-Model) that consists of multiple layers, which is set of input data onto set of suitable outputs and one or more hidden layers in between, in a directed graph with each layer completely connected to the subsequent one. ANN is one of the classification approaches used in data mining [11], it is the most suitable non-linear predictive model for rainfall prediction.

3.3.2. Support Vector Machine (SVM)

One of the supervised learning techniques which is commonly used for classification, regression, and ranking functions is Support Vector Machine. SVM is mostly used in classification problems for both linear and non-linear data [11]. SVM provides unique and optimal solution, the kernel function is selected based on the points of the variables in the hyperplane. The best separating hyper plane can be written as,

$W.X + b = 0$

Where w is a weight vector, the value of the attributes is referred as x, and b is scalar often referred as bias [11].

3.3.3. Naïve Bias

Naive Bayes Classifier is the simple and powerful supervised machine-learning algorithm used for predictive modeling. It considers all variables contribute in the direction of arrangement and they are equally connected [12]. The algorithm is based on a theorem called Bayesian Theorem and used when the coordination of the inputs is high, which assumes that features are statistically independent.

4. RESULT AND DISCUSSION

A total of 2157 instances with 10 attributes are collected to create a proper dataset for the design. The dataset is divided into two set of precipitation, Rain and No Rain. Based on the annual precipitation in Sulaymaniyah most of which occurring from November to April, the separation of the dataset is not equal; Table 2, mentioned in the previous section, is a summary of the obtained dataset. The Normalized dataset is tested in Weka using several algorithms and the performance and evaluation of each algorithm is tested.

4.1. Tested Algorithms

4.1.1. ANN

The classification of the dataset has been calculated based on a Feed-Forward Artificial Neural Network (FFANN) with back propagation gradient-decent algorithm. The design of FFANN is consist of three layers, input layer, one hidden layer, and output layer, the hidden and output layers are log- sigmoid. Several numbers of neurons in the hidden layer were selected to get the best classification performance. For this purpose, the dataset is divided in a percentage split of 60% training and 40% testing, and the optimal number of neurons in the hidden layer was chosen to be two.

The evaluation on the test split 40% dataset of total 863 instances. 756 instances are correctly classified and 107 instances are inaccurately classified with Mean Square Error (MSE) value of 0.2779. The total accurateness of the ANN design is 87.6014 %.

4.1.2. Support Vector Machine (SVM)

A Support Vector Machine with poly Kernel function is used for the classification of the dataset. Each data is plotted as a point in n – dimensional space, where n is the number of elements, with a particular coordinate value. The cross-validation test option with 10 folds is tested to get the performance. The evaluation performance on SVM with 10 folds cross-validation for the dataset. 1975 of the instances are correctly identified and 182 of the instances are incorrectly identified with Mean Square Error (MSE) 0.2905. The algorithm gives better performance compared with other optimization models with an accuracy of 91.5624 %.

4.1.3. Naïve Bayes

The Naïve Bayes uses Gaussian distribution for the assumption for each numerical attribute. The estimation of class condition probability is done by the classifier with the assumption that attributes are conditionally not dependent on each other.

Naïve Bayes algorithm uses 10 folds cross-validation test option to evaluate the performance on the dataset. The correctly identified instances are 1858, and 299 instances are incorrectly identified with the Mean Square Error (MSE) 0.3444. The algorithm gives a good performance of 86.1382 %. Although its accuracy is near to the one with ANN, but the test option is different for both algorithms.

4.2. Comparison on the Algorithms

Comparison of the output outcome with known class (pre- classified) data can be analyzed to check the performance of any supervised data mining techniques. Three evaluation parameters, Precision, Recall, and F measure, with the ROC area under the curve are used for the comparative analysis [5].

The confusion matrix is used to estimate the performance of the given algorithms. Performance is evaluated by estimating the Accuracy, sensitivity, and Specificity [13].

True Positive (TP): precipitation acceptably recognized as yes.

True Negative (TN): precipitation acceptably recognized as no.

False Negative (FN): precipitation inaccurately recognized as Yes.

False Positive (FP): precipitation inaccurately recognized as a No.

The confusion matrix on the precipitation is explained in the Table 3:

Table-3. The confusion Matrix for Rainfall prediction.

Real Value	Predictable Value	
	No	**Yes**
No	TN	FP
Yes	FN	TP

Source: Salih, et al. [13].

The estimation of the True Positive (TP) entities with respect to False Positive (FP) entities is calculated to get the Precision using the equation below:

$$Precision = \frac{TP}{(FP + TP)}$$

The estimation of the True Positive entities with respect to the (FN) False Negative entities is calculated to get the Recall using the equation below:

$$Recall = \frac{TP}{(TP + FN)}$$

Sometimes the precision and recall are not possible to choose the better algorithm for the performance evaluation. F- measure is the solution to this issue, which specify the average of recall and precision. It can be calculated as bellow:

$$F - Measure = \frac{Precision * Recall * 2}{(Precision + Recall)}$$

The proposed algorithms for the performance on the dataset are compared as giving in the Table 4. The result is expressively showing that the best algorithm for evaluating the performance on the rainfall is SVM with 10 folds cross-validation test option, which gives the performance of 91.5624 %. The accuracy of the algorithms is plotted clearly in the Figure 2.

Table-4. Comparative analysis for the proposed algorithms.

Algorithm	Accuracy	Precision	Recall	F- measure	ROC Area
ANN (Based on the 40% test split dataset)	87.6014 %	0.898	0.876	0.884	0.938
SVM	91.5624 %	0.912	0.916	0.913	0.800
Naïve Bayes	86.1382 %	0.906	0.861	0.875	0.941

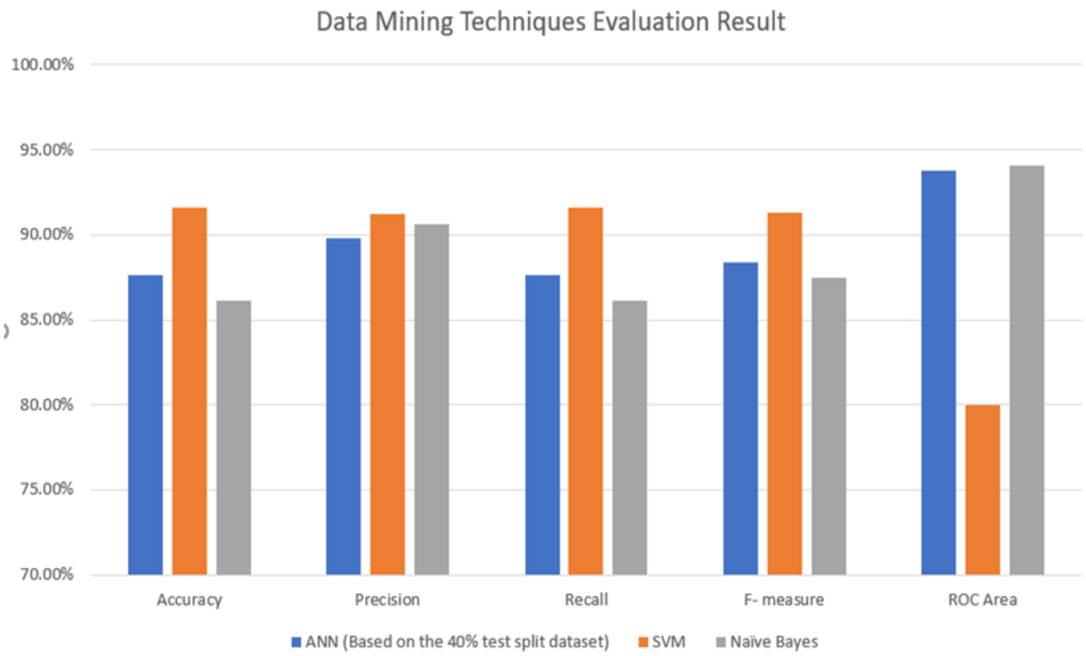

Figure-2. Comparison of Accuracy, Precision, Recall, F-measure, and Roc area among the three techniques.

5. CONCLUSION AND FUTURE WORK

This investigation achieved rainfall prediction in Sulaymaniyah city of Kurdistan region in Iraq. three data mining techniques used to perform the prediction of rainfall as Support Vector Machine, Naïve Bayes and Artificial neural Network. The data used for this work are collected from the weather forecast department in Sulaymaniyah city. The case data covers the period of 2013 to 2018. Three accuracy measures, precision, recall and f-measure are used to estimate the performance analysis of the data mining approaches and results are presented in tables and graphs. A classification framework is used for sufficient estimation, in which the input data went through a pre-processing stage, got cleaned and normalized before classification process. The selected data samples are transferred to a spreadsheet file for further processing to be suitable for data mining approaches. The data set were normalized to minimize the effect of scaling on the data and saved as a Commas Separated Value (CVS) file format. Conferring the results, used classification approaches performed well, but the best data mining technique to predict the rainfall is SVM, which gives the performance of 91.57 %. Further predictions can be performed for future work by examining more classification approaches and climatic features on different weather data in diverse parts in the district.

Funding: This study received no specific financial support.
Competing Interests: The authors declare that they have no competing interests.
Acknowledgement: Both authors contributed equally to the conception and design of the study.

REFERENCES

[1] N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay, "Development and analysis of artificial neural network models for rainfall prediction by using time-series data," *International Journal of Intelligent Systems and Applications*, vol. 10, pp. 16–23, 2018. Available at: 10.5815/ijisa.2018.01.03.

[2] D. Chauhan and J. Thakur, "Data mining techniques for weather prediction: A review," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 2, pp. 2184-2189, 2014. Available at: https://doi.org/10.22214/ijraset.2017.11353.

[3] F. Sheikh, S. Karthick, D. Malathi, J. S. Sudarsan, and C. Arun, "Analysis of data mining techniques for weather prediction," *Indian Journal of Science and Technology*, vol. 9, pp. 1-9, 2016.

[4] S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative analysis of data mining techniques for Malaysian rainfall prediction," *International Journal on Advanced Science, Engineering Information Technology*, vol. 6, pp. 1148-1153, 2016. Available at: https://doi.org/10.18517/ijaseit.6.6.1487.

[5] S. Aftab, M. Ahmad, N. Hameed, M. S. Bashir, I. Ali, and Z. Nawaz, "Rainfall prediction in Lahore City using data mining techniques," *International Journal of Advanced Computer Science and Applying*, vol. 9, pp. 254–260, 2018.

[6] V. B. Nikam and B. Meshram, "Modeling rainfall prediction using data mining method: A Bayesian approach," in *Proceeding International Conference on Computational Intelligence, Modelling and Simulation*, 2013, pp. 132-136.

[7] K. Abhishek, A. Kumar, R. Ranjan, and S. Kumar, "A rainfall prediction model using artificial neural network," in *Proceeding 2012 IEEE Control Syst. Grad. Res. Colloquium, ICSGRC 2012, no. Icsgrc*, 2012, pp. 82-87.

[8] R. V. Ramana, B. Krishna, S. Kumar, and N. Pandey, "Monthly rainfall prediction using wavelet neural network analysis," *Water Resources Management*, vol. 27, pp. 3697-3711, 2013. Available at: https://doi.org/10.1007/s11269-013-0374-4.

[9] J. Joseph and T. Ratheesh, "Rainfall prediction using data mining techniques," *International Journal of Computer Applications*, vol. 83, pp. 11-15, 2013. Available at: https://doi.org/10.5120/14467-2750.

[10] M. Narvekar and P. Fargose, "Daily weather forecasting using artificial neural network," *International Journal of Computer Applications*, vol. 121, pp. 9-13, 2015. Available at: https://doi.org/10.5120/21830-5088.

[11] R. Sukanya and K. Prabha, "Comparative analysis for prediction of rainfall using data mining techniques with artificial neural network," *International Journal of Computational Science and Engineering*, vol. 5, pp. 1–5, 2017.

[12] S. D. Jadhav and H. Channe, "Comparative study of K-NN, naive Bayes and decision tree classification techniques," *International Journal of Science and Research (IJSR)*, vol. 5, pp. 1842-1845, 2016. Available at: https://doi.org/10.21275/v5i1.nov153131.

[13] Y. M. M. Salih, A. Kattan, and T. Çevik, "Detection of motorway disorders by processing and classification of smartphone signals using artificial neural networks," *International Journal of Natural Sciences Research*, vol. 4, pp. 56-67, 2016. Available at: https://doi.org/10.18488/journal.63/2016.4.3/63.3.56.67.

Views and opinions expressed in this article are the views and opinions of the author(s), International Journal of Advances in Life Science and Technology shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.