ANALYSIS OF PRINCIPAL COMPONENTS AND CLASSIFICATION ENHANCEMENT FOR RAINFALL PREDICTION

K. Varada Rajkumar

Research Scholar, Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India

Dr. K. Subrahmanyam

Professor, Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

Abstract - Rainfall detection becomes an important task during varying climatic conditions. It is become very essential to examine the changing patterns of the rainfall and try to detect the rain. The rainfall databases are highly susceptible to noise, inconsistent and missing of data at present day in the real-world. So, the data obtained must be repaired by plugging absent values and eliminating the unrelated data. For this min-max normalization is employed for transferred the data into [0, 1] range. After that it was operated with z-score normalization, where the data values are distributed in a small scope. Then the feature extraction using PCA and ACO is functioned. Finally, naive bayes classification is applied to compare the performance of with and without pre-processing and the classes representation based on the different principal components (PC) of feature selection were calculated with and without pre-processing. Monthly rainfall data sets that are downloaded from Indian meteorological department (IMD). Monthly rainfall for years 1901 to 2019 are taken for analysis. The missing data and the outliers were detected every month. For investigation, the value distributions and features using PCA and ACO of rainfall feature before and after normalization is observed. The min-max normalization and z-score normalization concluded that they do not change the autocorrelation. The ACO based method found efficient when compared to PCA method. Classification error with and without data pre-processing is measured and found that the pre-processing rainfall data enhances the model performance.

Keywords: ACO, Rainfall Detection, Rainfall Prediction, Min-Max, Normalization, Classification, PCA, Pre-Processing

1. Introduction

For all the flora and fauna on planet earth, rainfall is a very significant aspect. It is important not only for human beings, but for animals, plants, and other living things as well. In agriculture and farming, it plays a critical role. It is one of the most valuable properties on the planet. The varying climatic conditions and rising carbon gases have contributed to the condition that the planet earth is unable to provide the requisite amount of rainfall and that the water available on earth is unable to fulfil human needs in their daily lives. It is therefore very interesting to study the evolving rainfall patterns and try to identify the rain not just for human needs, but also to deter any natural catastrophe that could arise due to unexpected heavy rainfall. So, it has become the focus of computer scientist and engineers to stay updated by being aware of the changing climatic conditions. The databases are highly susceptible to noise, inconsistent and missing of data at present day in the real-world. Databases are huge in size (often as gigabyte or higher than it) and originates from sources as multiple and heterogeneous. The mining results will be low when low quality of data used. Many numbers of data pre-processing techniques are used improve the accuracy and efficiency of database. To remove noise, correction of unpredictability in the data can done by data cleaning. Through using data aggregation, such as a data repository, data can be combined from different channels into a cohesive data store. Data reduction is used to reduce the size of the data by, for example, aggregating, removing redundant functionality, or clustering. When data is normalized to fit into a narrower range such as 0.0 to 1.0.0, data transformations such as normalization are applied. The accuracy and usefulness of a database can be enhanced by this methodology.

The remaining section of the article is structured as: Section 2 Literature Survey and Section 3 discusses the data Pre-processing. The feature selection demonstrated in Section 4. Section 5 on Classification and the results are discussed in Section 6. Lastly, the conclusion of the research is deliberated in Section 7.

2. Literature Survey

To boost the accuracy of a new generation of stochastic monthly rainfall prediction models tested by four separate pre-processing techniques, ebtehaj et.al., (2020) was proposed. But in certain situations, the presence of some of the deterministic terms may limit stochastic models. Transforms of normalization are often unable to adequately normalize the sequence. Ouyang, q. Introduced echo state networks (esn) and multi-gene genetic programming (mggp) to predict the monthly rainfall. And luy, and w (2018). However, the prediction capacity tests for longer lead times show that the best ssa-esn hybrid model can reasonably estimate rainfall at least 24 months ahead, however for lead times longer than that the model cannot guarantee reasonable precision. This is attributed to the accumulation of long lead times for predictive errors. As data pre-processing methods for the decomposition of raw rainfall data into subseries, the ensemble empirical mode decomposition (eemd), singular spectrum analysis (ssa) for long-term forecasting and wavelet transform (wt) for short-term forecasting were applied compared to support vector regression (svr) and the predictive efficiency of the models was evaluated on the basis of multiple parameters. The aim of this research was to establish a hybrid model for short-term runoff time series forecasting (wen-jing niu et.al., 2020). As a first, it used variational mode decomposition (vmd) to decompose the initial nonlinear natural streamflow into multiple subcomponents with varying frequencies and resolutions and the extreme learning machine (elm) as a second feature that excavates the dynamic relationship between input and output concealed in each subcomponent. Reasonable network parameters for each elm model were found using the sine cosine algorithm (sca). But it is still important to reliably capture the potential reservoir streamflow transition. In runoff forecasting for danjiangkou reservoir in china, multiple statistical assessment steps and feasibility were investigated in the hybrid process. In contrast to several control models, the result provided detailed streamflow prediction information for operators as a superior forecasting model.

Zou et.al. (2020) proposed four types of data pre-processing standardization methods and these methods are used for forecasting the displacement of bank landslides with different deformation mechanisms for comparison in the three gorges reservoir region. The results revealed that (1) the selected normalization method greatly affects the forecast efficiency (2) the selected forecast model was closely related to the method of normalization and the less reliant process of landslide deformation and (3) the optimal method is not the max-min normalization solution, and the zero-mean normalization method is optimal for the particle swarm opt for although the logarithmic normalization approach is suitable for the extreme learning machine method, it is important to carefully choose certain methods of data processing in landslide displacement forecasting. The artificial neural network (ann) model was introduced by simmonds et.al. (2017) and the record flow forecasting purposes were handled with the evolution conditions of the current hydrometric data and work phase related process.

Asadi et.al. (2013) suggested a hybrid smart runoff prediction model with a mixture of data pre-processing techniques, genetic algorithms and a levenberg-marquardt (lm) algorithm for learning to feed forward neural networks and data pre-processing techniques that maximize the model's accuracy. This approach was extended to the aghchai watershed to forecast runoff. The outcome shows that this method was able to more reliably forecast runoff than the model of the adaptive neuro fuzzy inference system (anfis) and the artificial neural network (ann) model. Wu & chau (2011) attempted to eradicate the lag impact of two aspects using modular artificial neural network (mann) and singular spectrum analysis (ssa) data pre-processing in the ann modelling phase two watersheds are explored in china .the findings suggested that mann did not demonstrate effective benefits over ann.at all three prediction horizons, the ann or mann with previous runoff only as model feedback is also established and contrasted with the ann (or mann) r-r model, resulting in the latter outperforming the former irrespective of being correlated with/without ssa. In the prediction model, the ssa dramatically improved efficiency and eliminated the lag effect. Coupled with ssa, the ann r-r model had better performance than other components. The assessment of rainfall time series using modular artificial neural networks combined with data pre-processing techniques was advised by wu et.al. (2010). This proposed model was contrasted with three benchmark models, namely artificial neural network (ann), linear regression (lr) and k-nearest-neighbours, modular artificial neural network (mann) (k-nn).three data pre-processing strategies, including singular spectrum analysis (ssa), principal component analysis (pca) and moving average (ma) under data pre-processing mode were combined for each of lr, k-nn and ann. Mann was then coupled with ssa and the result showed improved mann benefits over other models and it is possible to derive an efficient rainfall forecasting model from mann coupled with ssa.

Tian et.al.(2018) has suggested a hybrid forecasting method that addresses challenges in the learning process of wind speed forecasting in hybrid systems. 1) a modern data pre-processing algorithm built based on the proposed hybrid data pre-processing technique, where the benefits of each meritorious portion are combined to produce an effective method of data pre-processing. 2) the recently proposed multi-objective satin bowerbird optimizer algorithm has strengthened the elman neural network model, offering better predictive output precision and stability than other benchmark models. Ouyang et.al., suggested a self-adaptive data processing

technique known as ensemble empirical mode decomposition (eemd) (2016). this ensemble empirical mode decomposition (eemd) is used to preprocess a set of rainfall data and apply phase-space reconstruction methods to design the prediction model's input vectors. Help vector regression (svr) prior to eemd development. The case study was taken at the weather station in changchun, china. An auto-regressive integrated moving average model, a standard three-layer feed-forward artificial neural network model and a support vector regression model have shown the potential of the proposed hybrid model. Normalized mean squared error (nmse), correlation coefficient (cc) and nash-sutcliffe efficiency (nse) were tested for the predictive performance of these models based on mean absolute percent error (mape). The suggested hybrid model had the lowest values of 0.10 and 14.90 for nmse and mape, and the highest values of 0.91 and 0.83 for nse and cc.

3. Data Pre-processing

The data collected so far has been chaotic, with some missed values and some unnecessary data. The collected data must then be restored by filling in the missing values and deleting the obsolete information. In Real world of dataset, missing of data is common one and the missing of data will be in different ways.
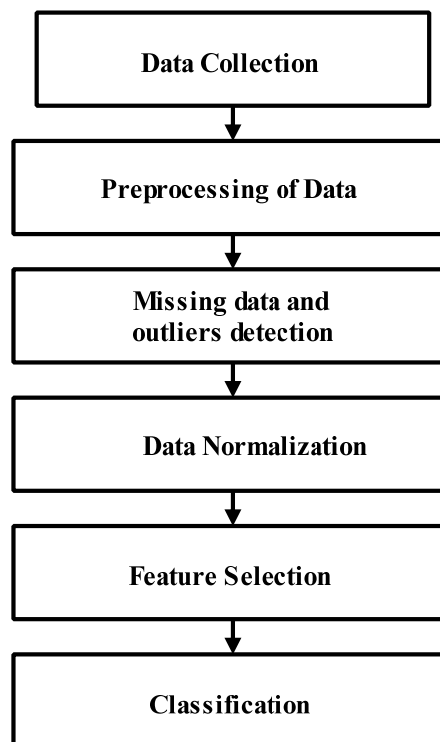

Fig.1. Flow chart of data processing

The Missing data will make sagacious effect on the result analysis with a conclusion of unreliable. It is necessary to learn and understand why the data is missing. The loss of data may be random or incomplete, but there is no association between the indicator and the effect, where the data sample always represents the population [11]. If a pattern lacks data, it relates to a response, leading to a significant model bias. Many methods have been proposed to deal with lost data [12, 13, 14], and they are categorized into two tactics. The first solution explicitly excludes the missing details. If the missing data has been spread at random or if it is missing, it may be linked to a predictor. The predictor will have zero correlation with the response and a large enough dataset, where this leads to little effect on the performance analysis when removing the missing data. The removing of missing data simply will be inappropriate when it does not satisfy with the above condition. Based on the rest of the results, the second approach was accomplished by filling in or attributing missing information. There are two ways of approaching this technique. In the first step, to fill in the missed value, the average predictor is used. In the second approach, the learning algorithm is used to predict the missing value, such as bayes or the decision tree [15]. It is documented by the imputation with extra ambiguity. Figure 1 displays the flow map of measures involved in data processing. For dataset, missing values and outliers should be detected

3.1 Missing values

In dataset a, the missing values was expressed by nan. So that numbers of missing values for each feature were calculated and defined with percent of missing values for each feature. The missing values are less than 10% for other features and then replaced with mean values for each feature.

3.2 Outliers

The outlier was defined as an observation point. The observation point was outlying from mainstream data. The data visualization is the most efficient way to identify outliers among other methods. A model's analysis ability can be broken when outliers at present. For example, if the data range is not specified it is hard to identify outliers because outliers will have a strong influence on regression fits and data scaling. By looking at a figure, some suspected observations can be pointed, and these values are checked whether they are scientifically valid (e.g. positive weight). Based on exact valid reasons, removal of outliers can do. An alternative way is used that transform data to minimize the effect caused by outliers instead of removing the outliers. Serneels proposed a spatial sign in 2006 which transforms the predictor values onto a new sphere [16]. The hampel function is used to detect the outliers for each feature when all the missing values are repaired. Since the datasets were time-based series, the slide windows are used to detect outliers with a window size to be 20. Figure.2 shows the number of outliers which are detected for each feature when window size of 20 set.

3.3 PCA Methods for Missing Data

A dataset with missing data should be handled with a process of complete analysis. By this process algorithm can implemented for a further work and this process is the one which has all analysis measurements and mean imputation. This process helps dataset so that missing element can be added by sample mean but this method undergoes to higher loss of data and terrible results. This method is a two-step procedure framed for a certain type of missingness and for implementing in multi-rate missingness. The multiple imputation lifting are methods used. The main aim is to evaluate method that integrates missing data handling and model building for pca. Consider, pca model has a complete data matrix but when they are put to test to detect principal components in the presence of missing data it was not in a complete manner. A least squares type of method used singular value decomposition based on that missing values are filled with sample mean proposed. The algorithm alternates between minimizing

$$C = \sum_{ij} \left(1 - M_{ij}\right)\left(X_{ij} - \sum_{k} t_{ik} P_{jk}\right)^2 \qquad (1)$$

had either fixed scores T or fixed loadings P where $M_{ij} = 1$ if X_{ij} is zero and missing otherwise. The First set of updated equations is

$$t_i^T = x_i^T A_i (A_i^T A_i)^{-1} \qquad (2)$$

Where, t_i is a ith column of T, x_i is a ith column of X, and A_i be d × a matrix with elements

$A_{jk} = (1 - M_{ij})p_{jk}$.

The second set of updated equations is

$$P_j^T = (B_j^T)^{-1} B_j^T x_j^T \qquad (3)$$

Where, p_j is a jth row of P, x_j is a jth row of X and B_j be n × a matrix with elements

$B_{ik} = t_{ik}(1 - M_{ij})$.

the model had additional loading vector with corresponding principal component equal to all ones for estimating address of μ. This model makes a resistance over the reconstruction error derivation in PCA problem and applies the changes in reconstruction error as a point of intersection.

3.4 Data Normalization

Scaling or a mapping technique or a pre-processing stage [1] is known as normalization. We may discover new ranges from an established range at that point. It adds more to the function of prediction or forecasting [2]. Normalization maintains greater variation in planning and forecasting, which leads to closer shifts in data observation. One of the steps of data pre-processing is data normalization. Normalization increased the performance and accuracy of prediction. The data normalization was determined with methods as followed below

3.5 Min-max and z-Score Normalization

The Normalization Technique of Min-Mix preserves the association between the original results. Min-Max standardization is a straightforward approach that precisely matches the data along with a pre-defined boundary inside a pre-specified boundary. Min-Mix Normalization [3-6] provides a linear translation to the original data set. For a Harmony and balanced dataset, The Data should be normalized between 0 and 1. where Equation. (4) was used to normalize the dataset.

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (4)$$

If x is the actual data, x max represents the maximum value of the original value of the attribute, and x min represents the minimum value of the original value of the attribute.

3.6 Z-Score Normalization

The normalization of the z-score is also known as zero-mean normalization. The values for an attribute, A, are normalized in terms of mean (i.e. mean) and standard deviation of A in this normalization. So that A value, v_i , of A was normalized to $v_i{'}$ by computing

$$v_i{'} = \frac{v_i - \bar{A}}{\sigma_A} \qquad (5)$$

where $\bar{A}$ and σA are the Mean and Standard Deviation of attribute A.

The Difference between min-max normalization and z-score normalization was that min-max normalization transfer data into values from 0 to 1. However, the min- max normalization, applied after had no longer input a feature value that fell outside of the original data range.

4. Feature Selection

The Feature Selection is often vital data processing step prior to applying a learning algorithm. For a Better understandable model with different visualization technique is possible by Reduction of attribute dimensionality. Reduction of attribute dimensionality is the process used for identifying and remove unrelated and dependable information as much as possible. The Min-Mix Normalization Technique preserves the relationship with the original The Learning Algorithms are run faster and more efficiently due to a decline in data dimensionality and can increase their precision later in future classification. It was observed that the minimal set of attributes resulting in the distribution of probabilities of data groups was as similar as possible to the initial distribution.

4.1 PCA

In 1901, Pearson introduced the first PCA, and Hotelling produced it independently in 1933. In data processing, this has been verified as an important methodology. The key concept is to minimize the dimensionality of a data set composed of several unified variables while preserving as considerably of the heterogeneity in the data set as possible. In the PCS method, all initial variables are used to produce a smaller number of main components (PCs) used to estimate the original variables. When PCs are uncorrelated and arranged, the first few maintain much of the heterogeneity present in the original collection. Consider a data matrix X with n (observation) rows, column p (variables). Let X's matrix of covariance be R, where $R = cov(X) = E(X^TX)$. The linear orthogonal transformed matrix Z is presented as

$$Z = XA \qquad (6)$$

Where Z is the PCs with elements of ith observation (i, j) and the main component of jth; A is a (p x p) matrix with covariance elements of X and has covariance elements of X. $A^T A = AA^T = I$.The Matrix of $X^T X$ was real and symmetric, so that it can be expressed as $X^TX = A{\wedge}A^T$, where K represents a diagonal matrix whose nonnegative entries are eigenvalues (λ_i; i =1,....,p) of X^TX.

The Total variance of the data matrix X is expressed as

$$Trace \left(\sum\right) = trace \left(A \wedge A^T\right) = trace \left(\wedge\right) = \sum_{i=1}^{p} \lambda_i \qquad (7)$$

On the other hand, the covariance matrix of principal components Z is expressed as

$$cov(Z) = E(Z^T Z) = E(A^T X^T XA) = \Lambda \qquad (8)$$

$$trace(Z) = trace(\wedge) = \sum_{i=1}^{p} \lambda_i \qquad (9)$$

Thus, after PCA transformation Z, the total variance of the data matrix X is equal to the total variance.

In the PCA solution, singular value decomposition (SVD) or covariance matrix causes of X are used. The SVD or covariance matrix determinants of X give their own values, λi; i = 1,...,p, to the eigenvectors A which represent the variance of each variable after the transformation of PCA. If eigenvalues are ordered by $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \lambda_p \geq 0$, where first few PCs can capture most of the variance of original data while other remaining PCs express mainly the noise in the data.

The Percentage of total variance defined by the first mth PCs is expressed as

$$V = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} . 100 \ \% \qquad (10)$$

To boost the data matrix properties that are retained, the Overall Data Variance (V) collection should be higher. A limited number of PCs are chosen for the sake of minimizing dimensionality, while much of the data variation in selected components persists. If the conversion avoids the collinearity of the regression variables, then the

number m factor selected in Eq(9).A higher overall variance, such as V= 95-99 percent, can be set for (9) (Hsu et al., 2002). The Original data matrix A can be reconstructed by a reverse operation of Equation. (5) as

$$X = ZA^T \qquad\qquad (11)$$

The original data can be filtered by choosing suitable m ($\leq$ p) PCs from Z and accompanying m eigenvectors from A.

4.2 ACO for Feature Selection

To execute ACO in such a manner that it can be suitable for unique solutions, the role selection task must be revamped. In ACO, it is important to graphically depict a problem. In ACO, features are represented by the nodes, and the edges represent the next feature's preference. The ant traversal approach performs the optimum search of the function subset through the graph in which the minimum number of nodes visited fulfils the traversal stopping criterion. In Fig.2 this is clearly represented. The ant has an option to select and add a function next to its path at the point a (dotted lines). It selects feature b, then c and then d based on the transition rule. The present subset {a; b; c; d} is calculated to fulfil the traversal stopping criterion after arriving at d (e.g. a high classification accuracy has been attained with the help of this subset). The ant has terminated its traversal for data reduction purposes and the feature subset of the outputs is given as a candidate.
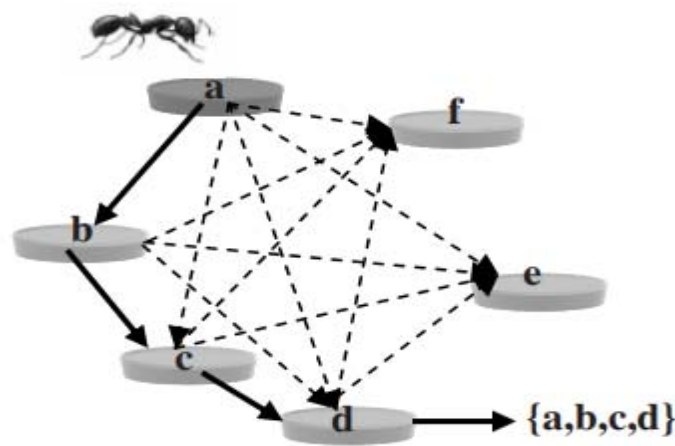

Fig. 2. ACO FS problem representation

Any sub-set evaluation function, such as an entropy-based metric, rough set dependency measure or the Fisher Discrimination Rate will be an effective functional implementation of traversing between characteristics (FDR). Combining the heuristic desire of traversal and edge pheromone speeds, the formation of the probabilistic transformation rule is completed, reflecting the chance of an ant at the feature I am selecting the feature j to move at the time t.

$$p_{ij}^k(t) = \begin{cases} \dfrac{[\tau_{ij}(t)]^\alpha \cdot [\eta_{ij}]^\beta}{\sum\limits_{I \in J_i^k}[\tau_{il}(t)]^\alpha \cdot [\eta_{il}]^\beta} & if \quad j \in J_i^k \\ 0 & otherwise \end{cases} \qquad (12)$$

Here k represents the number of ants, η_{ij} represents the heuristic desirability of selecting the feature from the feature i (η_{ij}) is optional, usually used for yielding a high algorithm performance , J_i^k represents the set of neighbour nodes i that are not visited by the and k. α ; 0, β ; 0 are the two parameter that estimates the relative importance of the pheromone value and heuristic information (the choice of α,β is estimated by experimentation) and (t) ij τ represents the amount of virtual pheromone on edge (i,j).

The entire process of feature selection in ACO is depicted in figure 3. The process of feature selection starts by assessing the number of ants k, and then placing it in the graph randomly (i.e. every ant starts with one random feature). In alternative the number of ants present on the graph and the number of features within the data are made equal. Every ant starts to construct a path at different feature. From their starting points they apply traverse edge probabilistic till they satisfy the traversal stopping criterion. The resulted subsets are then collected and calculated.
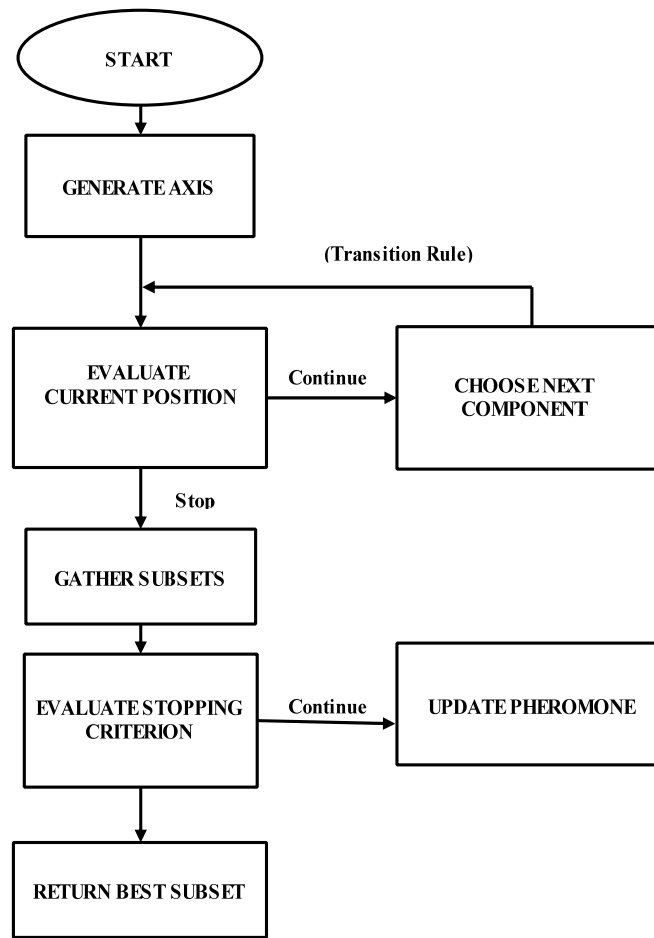
Fig. 3. Overview of ACO-based feature selection

The process halts if the optimal subset was found or if the algorithm has executed some specific number of times and the best feature subset of the output is encountered. If both the conditions hold, then pheromone updating is done and a new ant set is formed, and the above process carries on repeatedly. By use of the following formula the pheromone on each edge is updated.

$$\tau_{ij}(t+1) = (1-\rho).\tau_{ij}(t) + \rho.\Delta\tau_{ij}(t) \tag{13}$$

Where:

$$\Delta\tau_{ij}(t) = \sum_{k=1}^{n} \left(\gamma'(S^k) \middle/ \left| S^k \right| \right) \tag{14}$$

This happens when edge (i;j) has been traversed, $\Delta\tau_{ij}$ (t) is 0 otherwise. The value of $0 \leq \rho \leq 1$ denotes the decay constant employed to simulate the evaporation of pheromone; k S denotes the feature subset found by ant k. The pheromone is updated with respect to the measure of "goodness" of the feature subset of ant γ' and the size of subset. Hence by the above principal, the pheromone of the ants is updated.

"According to both the "goodness" measure of the subset γ of the ant's feature "and the size of the subset itself. By this interpretation, the pheromone is updated by all ants

5. Classification

It is a simple and efficient Naïve Bayes classifier. The Bayesian classifier is made up of a powerful assumption that has distinct features and is independent of each other (Langley et al., 1992). The higher posterior of the classification vector is chosen based on the classification of the given set of features. It is offensive against other advanced classifiers such as decision trees, while it is basic (Friedman et al., 1997). Furthermore, structure learning is not necessary to date, design and execution are simpler. The Naïve Bayes classifier paid considerable attention to addressing diverse classification concerns with excellent advantages, Friedman et al (1997). The Naïve Bayes classifier cannot fulfil the freedom expectation between features in real-world data due to these major limitations. Domingos and Pazzani (1997) found that with a heavy dependence on different characteristics, a naive Bayes classifier nevertheless worked better. If the freedom assumption is broken, it may not be optimal. The estimation accuracy of the Naïve Bayes classifier will be extremely susceptible to the correlated characteristics of real-world applications.

5.1 Basic model

The Naïve Bayes classifier is sometimes considered a basic Bayesian classifier. Bayes built it on a theorem. It is basic Bayesian Network (BN) and is especially appropriate for cases where inputs are of such high dimensionality. The Naïve Bayes classifier form is shown as a special BNN case in Figure 3.
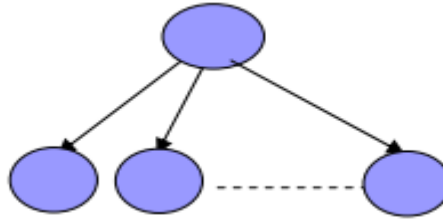

Fig. 4. Structure of naïve Bayes classifier

Consider, Set of samples x_1 , x_2 , , x_k with their associated class labels c_{x1} ,c_{x2} , ,c_{xk}, where $c_{xk} \in \Omega = \{c_1,c_2,c_L\}$ and assume further that samples have n features represented as $z_1,z_2,\ldots z_n$. By this task, usage of

samples can be learnt in a naïve Bayes model that predicts the label c_x for any future sample x . The BN classifier use Bayes rule to determine posterior of classification variable c based on the feature variables $z_1,z_2,\ldots z_n$, can be described as follows:

$$p(c \mid z_1, z_2, \ldots z_n) = \frac{p(z_1, z_2, \ldots z_n \mid c)\, p(c)}{p(z_1, z_2, \ldots z_n)} \qquad (15)$$

In application, Estimation of joint conditional probability $P(z_1,z_2,\ldots z_n|c)$ is not practical thing in Eq. (15). A standard practice has been generated that simplifies two amazing claims on Eq by the naive Bayes classifier. The first presumption is called the assumption of class-conditional freedom, i.e. all z1,z2,...zn characteristics are independent of each other from the classification variable c provided. It can be mathematically denoted as

$$p(z_1, z_2, \ldots z_n \mid c)\, p = p(z_1 \mid c)\, p(z_2 \mid c) \ldots p(z_n \mid c) \qquad (16)$$

The product of all marginal conditional probabilities is that which indicates a joint conditional probability. The second presumption is that the classification vector c explicitly relies on all characteristics z1,z2,...zn . If these two hypotheses are revealed on a general BN classifier, the naive Bayes classifier can be interpreted as:

$$p(c \mid z_1, z_2, \ldots z_n) = \frac{p(c)\prod_{i=1}^{n} p(z_i \mid c)}{p(z_1, z_2, \ldots z_n)} \qquad (17)$$

The model learning method was based on the training dataset given in the classification, which obtains the conditional likelihood of zi given c (i.e. p(zi | c)) and c prior (i.e. p(c)). Furthermore because p(z1,z2,...zn) is normal for a certain sample, it is overlooked in the process of classification. Therefore, the following model can be derived as:

$$c = \arg \max_{c \in \Omega} p(c)\prod_{i=1}^{n} p(z_i \mid c) \qquad (18)$$

By that prediction on class of each sample can be done. In application, Eq. (18) is often replaced by its logarithmic form as follows:

$$c = \arg \max_{c \in \Omega} \{\log(p(c)) + \sum_{i=1}^{n} \log(p(z_i \mid c))\} \qquad (19)$$

5.2 Data Set taken

The meteorological data that used in this research has been brought from Indian meteorological Department, based on previous 115 years data set calculation of Monthly Rainfall prediction was made in Andhra Pradesh, India.

6.   Results and Discussions

The experiments were performed in MATLAB platform using monthly rainfall data sets that are downloaded from Indian meteorological Department (IMD). Monthly rainfall for years 1901 to 2019 are taken for analysis. Fig 5 shows the Missing data detected every month. Fig 6 shows the Outliers detected every month.

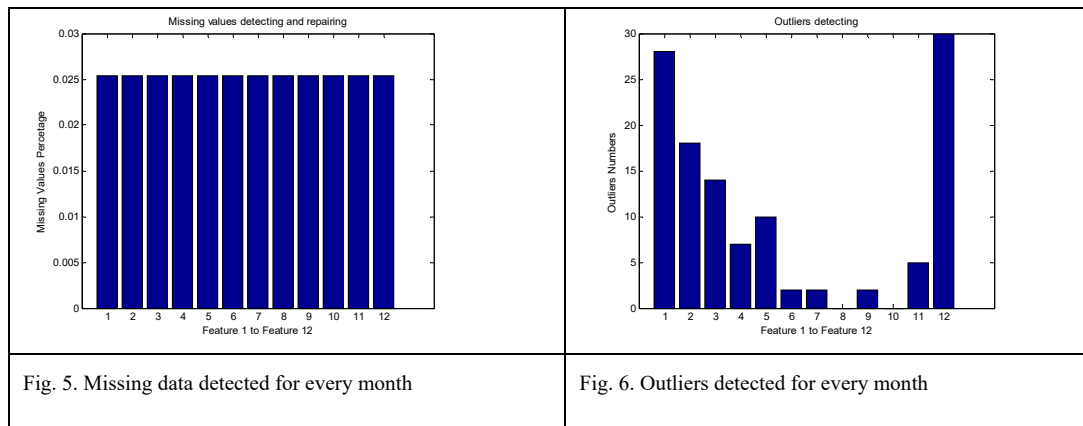| Fig. 5. Missing data detected for every month | Fig. 6. Outliers detected for every month |

Figure 5 showed the values distributions of April rainfall feature before and after normalization and figure 6 showed the values distributions of September feature before and after normalization.

Table 1. Performance Analysis of Missing data algorithms

Parameter	Mean Replacement	PCA using ALS
MSE	296.8912	40.4792
MAE	1.9168	0.7152
SSIM	0.9733	0.9939
DSSIM	0.0133	0.0031

Table 1 shows the performance analysis of missing data using mean-based replacement and PCA using ALS algorithm. PCA using ALS gives better result than mean replacement.

From figure 7 and figure 8 showed the original data which was not normalized distributed between large ranges. The data transferred to [0, 1] range when min-max normalization is applied. After that it was applied with z-score normalization, where the data values are distributed in a small scope. However, before and after normalization data had a similar distribution trend. Different normalization process is examined
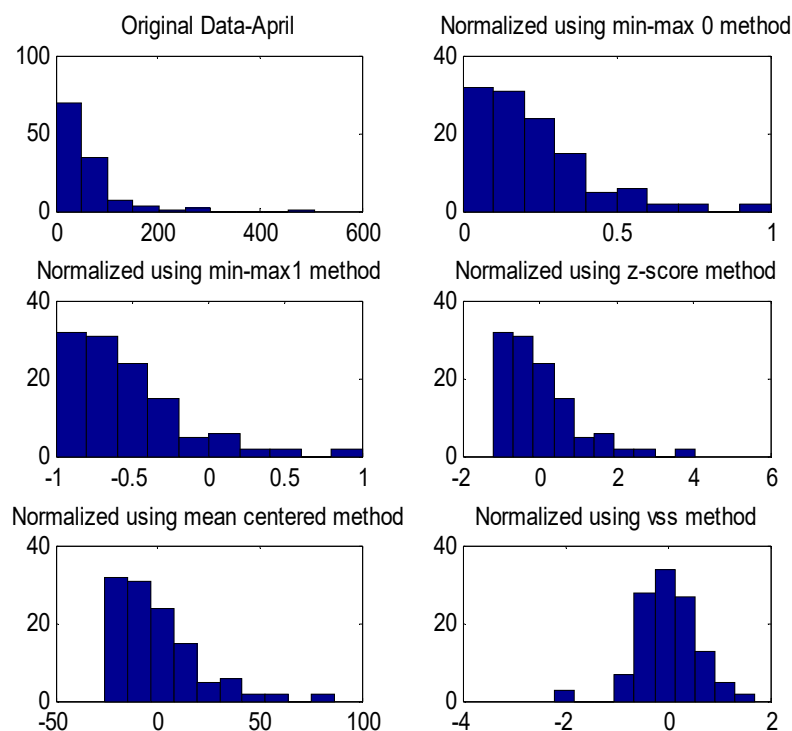

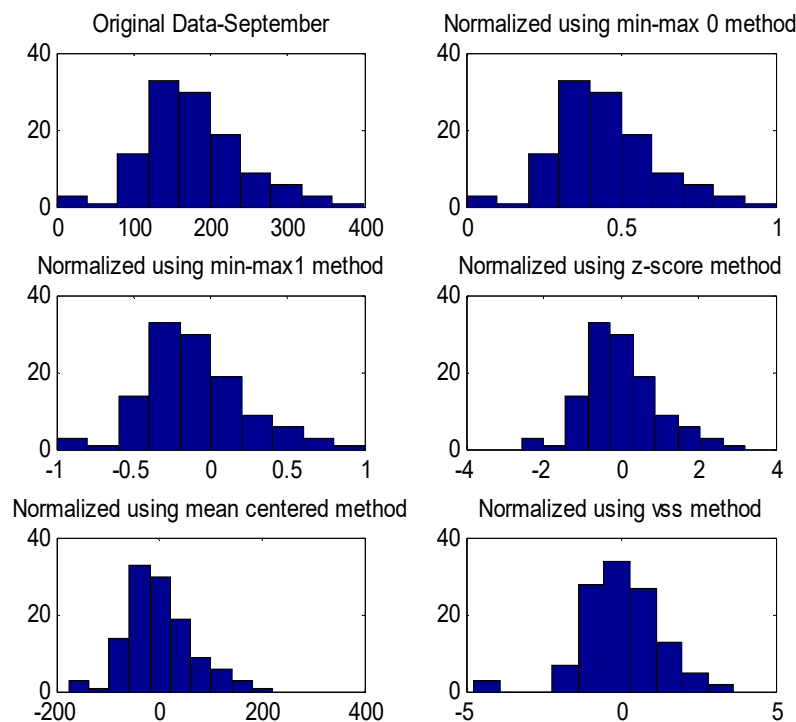Fig.7. Distributions of April rainfall before and after normalization

Fig.8. Distributions of September rainfall before and after normalization

The figure 9 and figure 10, shows features before and after normalization, that plot of auto-correlation line for #4 and #9. From these two figures it is known that they have the same auto-correlation lines before and after normalization. The Min-max normalization and z-score normalization concluded that they do not change the autocorrelation. But mean centered normalization slightly change the autocorrelation. Two variants of MMN methods that rescale data in the range of [0, 1] (MMN0) and [−1, 1] are considered (MMN1). The results of these approaches are similar, with fewer exceptions. The results are like datasets that have functions that are binary and real-valued. For binary-value characteristics, MMN0 scales the characteristics to 0 and 1, while MMN1 re scales the characteristics to -1 and +1.
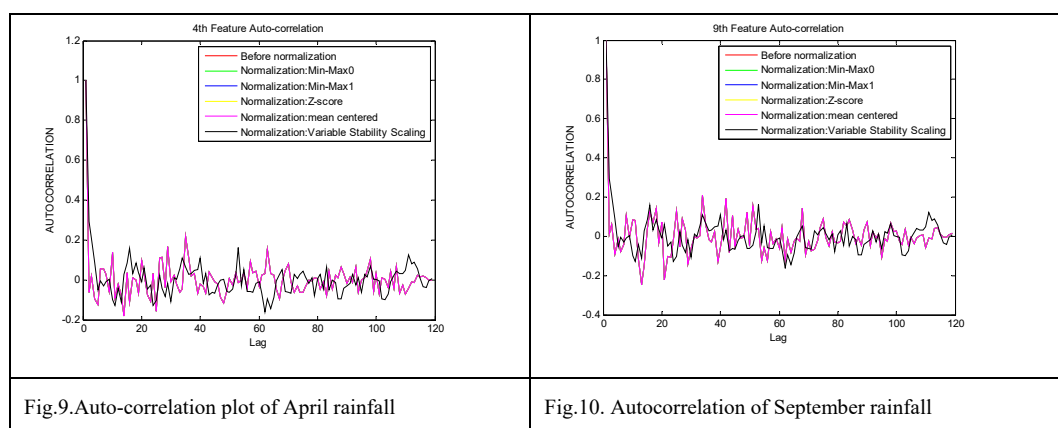

Fig.9.Auto-correlation plot of April rainfall	Fig.10. Autocorrelation of September rainfall

The latter approach only adjusts the proportions of the values that will be the same for all functions. With the MC technique, only the offset is removed from the information. Therefore, the issue of feature dominance, particularly if feature values have a higher numeric range, is not removed at all. Subtracting mean from this range does not affect the dominance of the feature, thus leading to accuracy of classification such as un-normalized data.

Fig 11 shows the Classes representation based on the 3rd and 4th principal components of feature selection. Fig 12 shows Classes representation based on the 9th and 12th principal components.

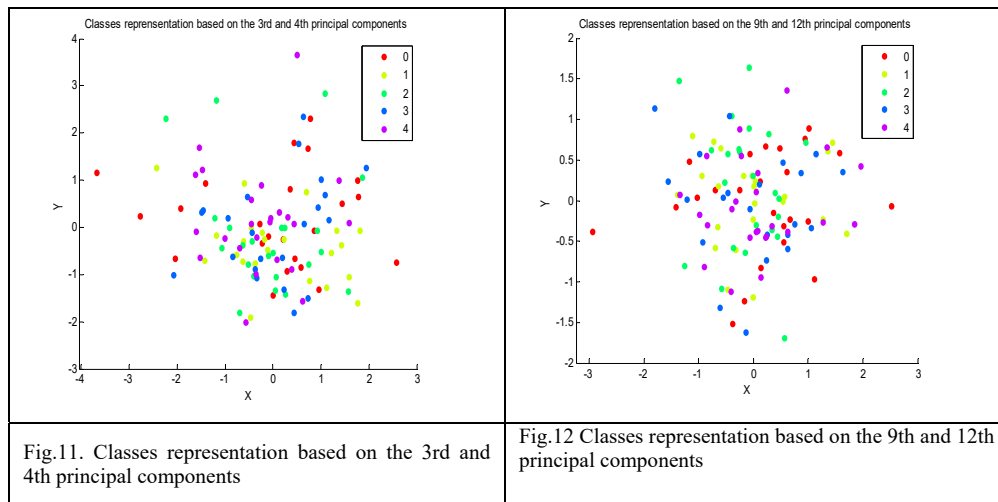| Fig.11. Classes representation based on the 3rd and 4th principal components	Fig.12 Classes representation based on the 9th and 12th principal components

Table 3 shows the performance of PCA based feature extraction for every month rainfall data. Table 2 shows the performance of ACO based method for feature extraction for every month rainfall data. Here ACO method gives better results but it takes more time to process.

Table 2. Feature Extraction by ACO based algorithm

Features	Mean	Standard Deviation	Entropy	RMS	Variance	Smoothness	Kurtosis	Skewness	IDM
JAN	0.2306	0.1814	3.5850	0.2887	0.0329	0.7345	1.9061	0.4977	0.1727
FEB	0.0938	0.2851	2.6175	0.2887	0.0813	0.5297	2.0160	0.6867	0.2915
MAR	0.0165	0.3010	1.9473	0.2887	0.0906	0.1653	2.5042	0.3280	0.2083
APRIL	0.0668	0.2933	2.4508	0.2887	0.0860	0.4451	2.6984	0.1283	0.6222
MAY	0.0537	0.2962	3.1887	0.2887	0.0878	0.3919	2.5399	-0.2287	0.4463
JUNE	0.0273	0.3002	2.6175	0.2887	0.0901	0.2465	2.4584	0.2262	0.0447
JULY	0.0600	0.2949	2.6175	0.2887	0.0870	0.4187	2.3354	0.0873	-0.1168
AUG	0.0043	0.3015	2.6175	0.2887	0.0909	0.0494	2.5210	0.3354	-0.3332
SEP	0.0655	0.2936	2.9183	0.2887	0.0862	0.4403	1.7277	-0.2308	0.5611
OCT	-0.0095	0.3013	2.2925	0.2887	0.0908	-0.1289	3.2414	0.4593	0.3335
NOV	0.0695	0.2926	2.9183	0.2887	0.0856	0.4548	1.9393	-0.0201	0.0606
DEC	-0.0123	0.3010	1.9473	0.2887	0.0907	-0.1733	2.5606	0.2787	0.0753

Table 3. Feature Extraction by PCA algorithm

Features	Mean	Standard Deviation	Entropy	RMS	Variance	Smoothness	Kurtosis	Skewness	IDM
JAN	0.0363	0.0605	2.4056	0.0672	0.0037	0.2248	2.3313	1.1431	0.0818
FEB	0.0628	0.0730	2.5000	0.0928	0.0053	0.3342	1.9297	0.7376	0.0783
MAR	0.3156	0.3518	3.0000	0.4560	0.1238	0.7163	1.3113	0.3427	0.4945
APRIL	0.2579	0.1834	3.0000	0.3097	0.0336	0.6735	2.0640	0.7057	0.6970
MAY	0.5643	0.6048	3.0000	0.7990	0.3658	0.8186	5.5530	2.0353	0.7138
JUNE	1.1106	0.4173	1.5488	1.1772	0.1741	0.8988	2.3332	0.0889	2.6073
JULY	2.0275	0.2690	0	2.0431	0.0724	0.9419	2.4761	-0.0452	4.0438
AUG	1.8484	0.4425	0	1.8942	0.1958	0.9367	2.5834	0.2680	3.7598
SEP	1.7845	0.4735	0	1.8387	0.2242	0.9345	2.0085	0.3751	3.6496
OCT	1.5229	0.9071	1.0613	1.7433	0.8228	0.9241	1.7087	0.2338	3.8327
NOV	0.4732	0.4313	3.0000	0.6219	0.1860	0.7911	1.1992	0.0608	1.1652
DEC	0.1153	0.1717	2.7500	0.1977	0.0295	0.4797	4.8549	1.7615	0.0435

Finally Naive Bayes classification is applied to compare the performance of with and without pre-processing. Figure 13 shows the Classification Error for various PCA components with and without data pre-processing. Result shows that pre-processing of data reduces the classification error. Hence pre-processing rainfall data enhances the model performance.
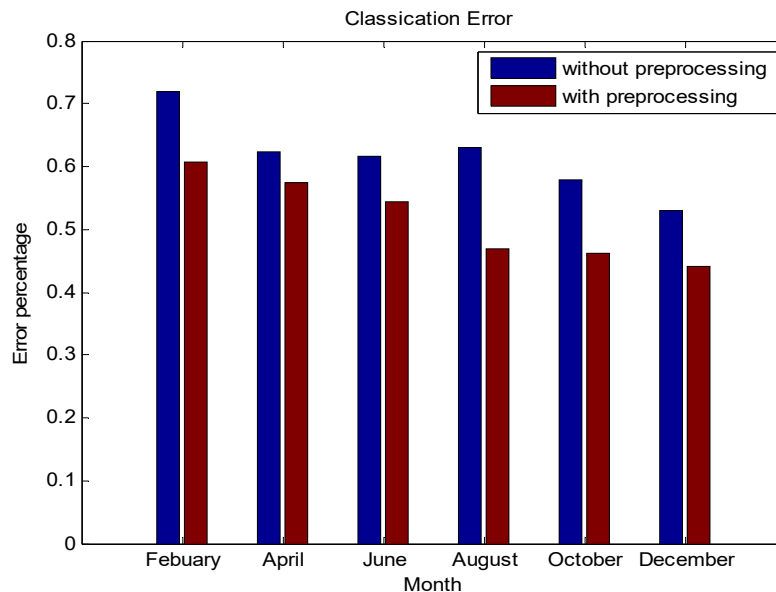

Fig.13. Classification error without and with Pre-processing

Table 4. Comparison table of Classification Error

Month	Without Pre-processing	With Pre-processing
February	0.7192	0.6076
April	0.6237	0.5740
June	0.6162	0.5450
August	0.6296	0.4699
October	0.5793	0.4614
December	0.5302	0.4405

7. Conclusion

This paper presents the rainfall detection and prediction using classified methods. The classification is investigated using the principal component analysis and pre-processing techniques. The noisy and missing data are corrected using the min-max normalization process. The data distribution was done in uniform using the z-score normalization method. The classification and class representation were carried out using the naive bayes classifier and PCA respectively with and without pre-processing. Database was collected from Indian meteorological department (IMD). The performance of PCA based feature extraction for every month rainfall data is observed. The performance of ACO based method for feature extraction for every month rainfall data shows its better nature compared to PCA but it takes more time to process. From various analysis and plots of features with autocorrelation it is found that the pre-processing enhances the classification method.

References

[1] I.Ebtehaj, H. Bonakdari, M. Zeynoddin, B.Gharbaghi, ''Evaluation of pre-processing techniques for improving the accuracy of stochastic rainfall forecast models,''in International Journal of Science and Technology, vol.17,pp. 505–524, 2019.

[2] Q.Ouyang and W.Lu,''Monthly Rainfall Forecasting Using Echo State Networks Coupled with Data Pre-processing Methods,'' in water Resource Manage, vol.32, pp.659–674,2018.

[3] Wen-jing Niu, Zhong-kai Feng, Wen-fa Yang, Jun Zhang, ''Short-term streamflow time series prediction model by machine learning tool based on a data Processing technique and Swarm intelligence algorithm,'' in Hydrological Science Journal,2020.

[4] Z. Zou, Y.Yang, Z. Fan, H.Tang, M.Zou, X.Hu, C.Xiong''Suitability of data pre-processing methods for landslide displacement forecasting,'' in Stochastic Environmental Research Risk Assessment, vol.34,pp. 1105–1119,2020.

[5] J.Simmonds , J.A Gómez ,A. Ledezma, ''Data Pre-processing to Enhance Flow Forecasting in a Tropical River Basin,''In: Boracchi G., Iliadis L., Jayne C., Likas A,(eds) Engineering Applications of Neural Networks, EANN 2017. Communications in Computer and Information Science, vol 744, pp.429-440, Springer, Cham, 2017.

[6]   S.Asadi, J.Shahrabi, P.Abbaszadeh, S. Tabanmehr, ''A new hybrid artificial neural networks for rainfall–runoff process modelling,' in Neurocomputing,vol.121,pp.470480,2013.

[7]   C.L.Wu, K.W Chau, ''Rainfall–runoff modelling using artificial neural network coupled with singular spectrum analysis,'' in Journal of Hydrology, vol. 399(3-4), pp.394–409, 2011.

[8]   C. L.Wu, K.W.Chau, C. Fan, ''Prediction of rainfall time series using modular artificial neural networks coupled with data-pre-processing techniques, 'in Journal of Hydrology, July 2010.

[9]   C.Tian, Y. Hao, J. Hu, J, ''A novel wind speed forecasting system based on hybrid data pre-processing and multi-objective optimization," in Applied Energy, vol. 231, pp. 301–319, 2018.

[10]   Q.Ouyang, W. Lu, X. Xin, Y. Zhang, T. Yu,''Monthly Rainfall Forecasting Using EEMD-SVR Based on Phase-Space Reconstruction, "Water Resource Manage, vol.30, pp. 2311–2325, march2016.

[11]   Y.Alimohammadlou, A.Najafi, C.Gokceoglu, ''Estimation of rainfall-induced landslides using ANN and fuzzy clustering methods: a case study in Saeen slope AzerbaijanProvince,Iran,''Catena,vol.120,pp.149-162,2014.

[12]   M.REN, B. WANG, Q. LIANG, G. FU, ''Classified real-time flood forecasting by coupling fuzzy clustering and neural network,'' International Journal of Sediment Research, vol.25, no.2, pp. 134–148, 2010.

[13]   A.K.Lohani, N.K.Goel, K.K.S.Bhatia,'' Improving real time flood forecasting using fuzzy inference system,'' Journal of Hydrology, vol.509, pp.25–41,2014.

[14]   L.H. Son, P.H. Thong, ''Some novel hybrids forecast methods based on picture fuzzy clustering for weather nowcasting from satellite image sequences,'' Applied Intelligence, vol.46, no.2, pp. 1–15,2017.

[15]   P.Singh, ''Rainfall and financial forecasting using fuzzy time series and neural networks based model,'' International Journal of Machine Learning and Cybernetics, vol.9, pp. 491–506, 2018.

[16]   S. Lee, J. Won, S.W. Jeon, I.Park, M.J.Lee , ''Spatial Landslide Hazard Prediction Using Rainfall Probability and a Logistic Regression Model,'' Mathmatical Geoscince, vol.47,pp.565–589,2015.

[17]   J. Wu, J. Long, M. Liu, ''Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm,'' Neurocomputing, vol.148, pp.136–142, jan2015.

[18]   A.M. Bagirov, A. Mahmood, A. Barton, ''Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach,'' Atmospheric Research, vol. 188, pp. 20–29,2017.

[19]   A.Danandeh Mehr, V. Nourani, V. Karimi Khosrowshahi, M.A. Ghorbani, '' A hybrid support vector regression–firefly model for monthly rainfall forecasting,'' International Journal of Environmental Science and Technology,Vol.16,pp. 335–346,2019.

[20]   Ro'fah Nur Rachmawatia, Iwa Sungkawaa, Anita Rahayua, "Extreme Rainfall Prediction using Bayesian Quantile Regression in Statistical Downscaling Modelling", ProcediaComputerScience, vol.157, pp.406-413,2019.

[21]   S. Cramer, M. Kampouridis, A.A. Freitas, A.K. Alexandridis, ''An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives,'' Expert Systems with Applications, vol.85, pp.169–181,2017.

[22]   J.George, L. Janaki, J. Parameswaran Gomathy, ''Statistical Downscaling Using Local Polynomial Regression for Rainfall Predictions – A Case Study,'' Water Resource Management, vol.30, pp.183–193, 2016.

[23]   X. Zhang, S. N. Mohanty, A. K. Parida, S. K. Pani, B. Dong and X. Cheng, "Annual and Non-Monsoon Rainfall Prediction Modelling Using SVR-MLP: An Empirical Study From Odisha,"IEEE Access, vol.8,pp.30223-30233,2020.

[24]   R.C.Deo, O. Kisi, V.P Singh, 'Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machineandM5Tree ode,''AtmosphericResearch,vol.184,pp.149175,2017.

[25]   A. Kusiak, X. Wei, A. P. Verma and E. Roz, "Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach," in IEEE Transactions on Geoscience and Remote Sensing, vol.51, no.4, pp.2337-2342,2013.

[26]   S. Bhomia, N. Jaiswal, C. M. Kishtawal and R. Kumar, "Multi model Prediction of Monsoon Rain Using Dynamical Model Selection," in IEEE Transactions on Geoscience and Remote Sensing, vol.54, no.5,pp.2911-2917,2016.

[27]   F. S. Marzano, G. Rivolta, E. Coppola, B. Tomassetti and M. Verdecchia, "Rainfall Nowcasting From Multisatellite Passive-Sensor Images Using a Recurrent Neural Network," in IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 11, pp. 3800-3812,2007.

[28]   J. Pucheta, C. Rodriguez Rivero, M. Herrera, C. Salas and V. Sauchelli, "Rainfall Forecasting Using Sub Sampling Nonparametric Methods," in IEEE Latin America Transactions, vol. 11, no. 1, pp. 646-650, 2013.

[29]   A. Haidar and B. Verma, "Monthly Rainfall Forecasting Using One-Dimensional Deep Convolutional Neural Network," in IEEE Access, vol. 6, pp. 69053-69063, 2018.

[30]   P. Zhang, Y. Jia, J. Gao, W. Song and H. Leung, "Short-Term Rainfall Forecasting Using Multi-Layer Perceptron," in IEEE Transactions on Big Data, vol. 6, no. 1, pp. 93-106, 2020.

[31]   C. Rodriguez Rivero, J. Pucheta, M. Herrera, V. Sauchelli and S. Laboret, "Time Series Forecasting Using Bayesian Method: Application to Cumulative Rainfall," in IEEE Latin America Transactions, vol.11, no.1, pp.359-364, 2013.

[32]   S.M.Hosseini,N.Mahjouri,'' Integrating Support Vector Regression and a geomorphologic Artificial Neural Network for daily rainfall-runoff modelling,'' Applied Soft Computing. 38,pp.329–345,2016.

[33]   D C R. Novitasari, H. Rohayani, Suwanto, Arnita, Rico, R Junaidi, Rr D N. Setyowati,R. Pramulya and F.Setiawan , ''Weather Parameters Forecasting as Variables for Rainfall Prediction using Adaptive Neuro Fuzzy Inference System (ANFIS) and Support Vector Regression (SVR)", Journal of Physics: Conference Series, Volume 1501, vol.2,2019, PP.1-11.

[34]   S.Georganos, A.M Abdi, D.E Tenenbaum, S. Kalogirou, ''Examining the NDVI-rainfall relationship in the semi-arid Sahel using geographically weighted regression "Journal of Arid Environments,vol.146,pp.64–74,2017.

[35]   M. I. Khan,R. Maity, "Hybrid Deep Learning Approach for Multi-Step-Ahead Daily Rainfall Prediction Using GCM Simulations," in IEEE Access, vol. 8, pp. 52774-52784, 2020.

Author Profile

K VaradaRajkumar is presently working as Assistant Professor of CSE at Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh. He received Master of Technology in Computer Science & Engineering during 2008-10 from JNT University Kakinada. Presently he is pursuing Ph.D. in the field of Machine Learning from K L University, Vijayawada, Andhra Pradesh.

Dr. K. Subrahmanyam Presently working as Professor of the Department of Computer Science & Engineering, K L University, Vijayawada, Andhra Pradesh, India. He received Ph.D from Andhra University. He has several papers in National and International Journals. His Research interest includes Software Engineering, Bioinformatics and Machine Learning.