IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-16, May 2019 E-ISSN: 2347-2693

A Rotation Forest Algorithm for Predicting BOD in River Water

J.A. Mangai[1*], B.B. Gulyani[2], R. Khanam [3]

[1,3]Department of Computer Science & Engg., Presidency University, Bengaluru, India
[2]Department of Computer Science & Engg., BITS-Dubai Campus, Dubai

Corresponding Author: alamelu.magai@presidencyuniversity.in,

DOI: https://doi.org/10.26438/ijcse/v7si16.17 | Available online at: www.ijcseonline.org

Abstract— Biochemical oxygen demand (BOD) is an important parameter for measuring the water quality especially the extent of water pollution due to organic compounds. The standard test for BOD requires a time period of 5 days with stringent conditions to be observed with regards to temperature, nutrients available and the lighting conditions suitable for the microbial growth. In order to predict BOD of river water in a cost-effective and efficient manner, in this paper a data driven ensemble method namely a Rotation Forest (RF) has been implemented. The model uses model trees M5 as base learners and hence the name rotation forest. Each base learner is trained using the rotated feature axes built on feature subsets computed using Principal Component Analysis (PCA). This helps to improve diversity in training the base learners and hence improves the predictive accuracy. Experimental analysis on available data sets shows that the correlation coefficient of a proposed approach is 0.9386 and RMSE of 0.5388. The predictive accuracy of this model is also compared with Multilayer Perceptron (MLP) neural networks model. However the proposed model has high correlation coefficient and low RMSE than MLP.

Keywords—BOD, rotation forest, ensemble,M5,MLP, PCA,Correlation Coeffecient,RMSE.

I. INTRODUCTION

Besides being an indispensable requirement for life on earth, water has become an important resource to sustain the life of all living organism. Therefore, in order to keep the water quality within a permissible limit according to the pollution control board there is a need to monitor the quality of water on a regular basis. Biochemical oxygen demand (BOD) represents the measure of the amount of oxygen required by microorganisms to decompose the organic matter present in the waste water under aerobic conditions. Thus BOD test is used to estimate pollution level of any kind of waste water either industrial wastes or domestic waste. According to the pollution control board, prior to discharge the waste in the natural water course there is an extreme need to treat the water. To design the appropriate size on water treatment plant BOD test is done, and measures the efficiency of waste treatment plant and track compliance waste water of discharge limit.The BOD5 test procedure involves measurement of the consumption of the dissolved oxygen needed by microorganism to break down the organic matter in present in the water sample in extreme aerobic condition [1]. By incubating a small number of water samples in BOD bottle saturated with oxygen and containing microbial seed and nutrients. The standard method for testing BOD is carried out in 5 days at a temperature of 20 degree c. BOD tests are time consuming and require strict control over the test conditions in order to achieve accurate results. Conditions that are conducive to biological growth need to be maintained while ensuring minimum interference from other factors such as oxidation of nitrogenous compounds by large population of nitrifying bacteria [2]. BOD test mimics the natural degradation of organic content in an ecosystem and is an important water quality indicator for organic water content. Hence, some predictive models are needed to provide for cost-effective and efficient BOD measurement and indication. Data mining models have been known to be good prediction tools.

Data Mining is an active research area with its roots from statistics, artificial intelligence, machine learning and data bases. It helps to discover previously unknown knowledge from huge data repositories. Its applications span almost all areas including chemical, medical, web mining, fault diagnosis, finance and so on. Data mining tasks are broadly segregated into descriptive and predictive analysis. Predictive tasks identify a model for the output variable in terms of the input variables in the data set. The data type of the output variable is numeric for regression and discrete for classification tasks. The dataset used in study is generally segregated into two sets training sets and test sets. The training set is used for training purpose of model and the test set is required to validate its performance.

To improve the predictive accuracy, the predictions made by multiple regressors/classifiers are aggregated. These techniques are called Ensemble methods. An ensemble method builds several base regressors/classifiers from the training data. It performs prediction on a test data by combining the predictions made by all base learners using majority voting [3]. The predictive accuracy of an ensemble model is high if the base learners are independent of each other and are better than a learner that does random guessing. The base learners are modeled using different techniques such as i) by manipulating the training set namely Bagging, Boosting and Random forests, ii) by manipulating the input features namely Rotation Forest and Random Subspaces, iii) by manipulating the class labels namely Error-Correcting output coding, or iv) by manipulating the learning algorithm. In this paper a rotation forest with model trees as base regressors for predicting BOD in river water is proposed.

II. RELATED WORK

Different attempts have been made successfully for the prediction of BOD using process based approaches. Number of deterministic models have been developed to measure quality of water in recent years, such as, QUAL2E, WATEVAL, QUAL2E ,MIKE11, WASP and HEC5Q [4]. The classical process-based modeling methods results in good predictions, but it has got certain limitation like involvement of much data calibration process. In addition to this, it depends on the approximation of various underlying processes. They cannot therefore be applied to situations beyond the hypotheses on which the model developed was based. In addition, too many model parameters make model computation intensive and slow. The main problem with process-based modeling approaches are primarily limited on water quality and high cost of monitoring quality of water. In the upcoming year, an alpha-trimmed ARIMA model for predicting BOD was proposed by [5]. Review of paper on the use of ANN in predicting and forecasting variable in water resource [6]. A neural network based model for predicting Dissolved oxygen from BOD and COD was proposed by [7]. An adaptive neuro fuzzy inference system (ANFIS) to predict BOD using 10 water quality parameters is compared with a similar model using only four parameters namely hardness, alkalinity, pH and DO as they were found to have significant relationship with BOD [8]. The model trained with all parameters was found to have better performance than the one trained using four parameters. ANN performance depends on many parameters such as network architecture, modeling parameters and training techniques in addition to pre-processing. Data driven models are a good alternative because, compared to deterministic model, they require fewer input parameters and input conditions [9]. Data mining applications are surveyed in water quality management [10]. A spatio-temporal analysis of coastal water quality using box plots in data mining and multi-variate statistical analysis using factor analysis was presented by [11]. A model based approach for predicting BOD in settled sewage is reported by [12]. Correlations between TOC with each one of BOD and COD for industrial waste water were proposed [13] using regression analysis and neural networks [14]. Much of the related research literature about using data driven models for BOD prediction have used a single base learner or regressor. However ensemble learners have been proved to improve the predictive accuracy by combining the predictions made by multiple learners [3]. These ensemble methods are found to reduce error and overfitting of data. Rotation forest is a new classifier ensemble method [15] that can be used to create an ensemble of classifiers or regressors depending on the base learners. There exists a detailed study of these ensembles for various applications [16][17][18]. Motivated by these facts an ensemble method namely rotation forest with model trees as base regressors is proposed for BOD prediction in this paper predictions, but the disadvantage is they involve much data calibration process. They also depend on different underlying processes being approximated. They cannot therefore be applied to situations beyond the hypotheses on which the model developed was based. In addition, too many model parameters make the model computation-intensive and slow. The major problems with process-based modeling approaches are limited data on water quality and the high costs of monitoring water quality.

III. METHODOLOGY

The data set used to train the BOD prediction model is in structured format as illustrated in Figure 1. In this figure rows represent the water samples and the columns indicate the chemical parameters. From a data mining perspective rows are known as instances and columns as features. The last feature is the output parameter to be predicted namely BOD. If the number of parameters used to train the model is M, then each water sample can be represented as a single point in a M-dimensional vector. All the N water samples are represented using this vector space model. The following sections 3.1, 3.2 and 3.3 have a description of the algorithms used in the proposed BOD prediction model.

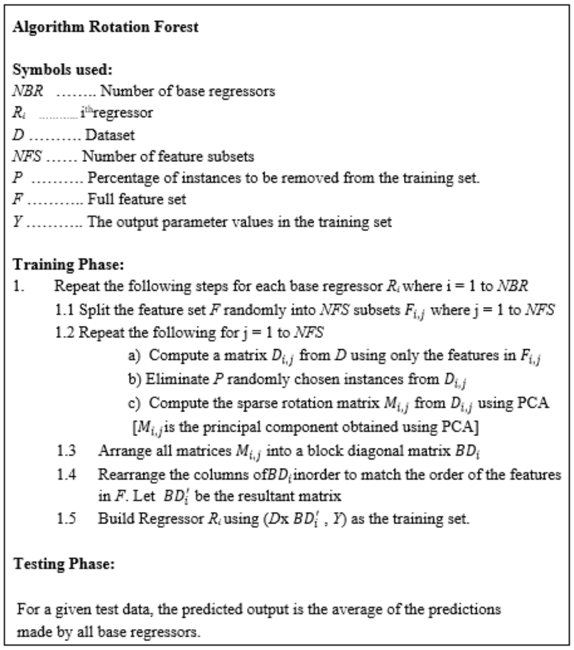
A. The Rotation Forest Algorithm

Ensemble prediction is a popular research area. Numerous experimental studies have proved that the practice of combining multiple learners has significantly improved the predictive accuracy. Bagging, Boosting and Rotation based approaches are the most common ensemble methods. Diversity is an important concept in ensemble theory. The rotation forest ensemble creates training sets for the base regressors from the original data set using PCA to rotate the original feature axes. For each base regressor the algorithm divides the feature space into a user defined number of subsets. A rotation matrix is constructed from these subsets using PCA as illustrated in Figure 2, later it is used to convert the test and the training datasets.

Features	F_1	F_2	F_3	F_4	F_5	…	…	F_M	**BOD**
Water Samples									
Sample 1									
………..									
Sample N									

Figure1. Training Data in Structured Format

It has been demonstrated that Rotation Forests is more accurate than Bagging, Boosting when size of dataset is large [14]. Rotation forests perform better due to the following reasons (i) PCA based feature selection on training data creates a diversity within the ensemble 2) decision trees which are more sensitive when the training data varies are used for base classification and (3) retaining all features extracted from the training data favors predictive accuracy of the base classifiers. The detailed pseudo code of the Rotation Forest algorithm for classification can be found in [15]. This paper proposes the use of rotation forest for predicting BOD in river water. The ensembles are constructed by manipulating the instances and features. The following section is a description of the base regressor used in the prediction model namely Model Trees.



```
Algorithm Rotation Forest

Symbols used:
NBR ........ Number of base regressors
R_i ............ i^th regressor
D .......... Dataset
NFS ...... Number of feature subsets
P .......... Percentage of instances to be removed from the training set.
F ........... Full feature set
Y ........... The output parameter values in the training set

Training Phase:
1.  Repeat the following steps for each base regressor $R_i$ where i = 1 to NBR
    1.1 Split the feature set F randomly into NFS subsets $F_{i,j}$ where j = 1 to NFS
    1.2 Repeat the following for j = 1 to NFS
            a) Compute a matrix $D_{i,j}$ from D using only the features in $F_{i,j}$
            b) Eliminate P randomly chosen instances from $D_{i,j}$
            c) Compute the sparse rotation matrix $M_{i,j}$ from $D_{i,j}$ using PCA
             [$M_{i,j}$is the principal component obtained using PCA]
    1.3 Arrange all matrices $M_{i,j}$ into a block diagonal matrix $BD_i$
    1.4 Rearrange the columns of$BD_i$inorder to match the order of the features
        in F. Let $BD_i^{\prime}$ be the resultant matrix
    1.5 Build Regressor $R_i$ using (Dx $BD_i^{\prime}$ , Y) as the training set.

Testing Phase:

For a given test data, the predicted output is the average of the predictions
made by all base regressors.
```

Figure2. The Rotation Forest Algorithm

B. *Model Tree*

The base regressors in the proposed model for predicting BOD are Model Trees. These are alternative to statistical regression. Model Trees are formed by combining regression trees with linear regression. Regression trees are decision trees but the leaf nodes of a regression tree are numeric rather than categorical values. The Regression tree are built using Recursive partitioning algorithms [22]. This algorithm repetitively partitions a subset of training samples into smaller subsets until a certain condition of stoppage is fulfilled. To determine the best splits at each node. As illustrated in Eq1. The parameter that minimize the least square error are often used to build a regression tree.

$$\text{Least Square error} = \frac{1}{n}\sum_{i=1}^{n}(a_i - p_i)^2 \quad (1)$$

Where n is the number of training examples, a_i and p_i are the actual and predicted result of the ith example. The value of a leaf node l as illustrated in Eqn.2 is the average target values y_i of all the examples that reach this node.

$$\text{value}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} y_l \qquad (2)$$

Where n_l in the leaf node l is the number of examples. The error rate at a leaf node l is calculated as,

$$\text{Error (l)} = \frac{1}{n_l}\sum_{i=1}^{n_l}(y_i - \text{value}_l)^2 \qquad (3)$$

The splitting criteria choses the split that reduces the error of the resulting tree after the split. For nonlinear data modeling regression trees are more accurate but it is sometimes difficult to understand [23].Therefore, to form the model tree along with the regression trees linear regression is also combined. For model tree construction, M5 trees [24] the first step is to determine the standard deviation of the target values of the examples T that reach a node. The set T cab be labeled as a leaf node or it can be further divided into subsets based on the test result. With each subset the process repeats until T has few examples or the values in T slightly varies. Eqn.4 gives the expected error reduction after dividing T into 'p' number of partitions based on a test condition.

$$\Delta\text{error} = \text{S.D.}(T) - \sum_{i=1}^{p} P(i)\text{x S.D}_i \qquad (4)$$

Where S.D. is the standard deviation. M5 chooses one splits such a way that it maximizes reduction in error among the set of candidate splits. Using standard regression techniques multivariate linear regression models are built at every node.

The model is also simplified using pruning and smoothing techniques. The following section is a description of the PCA algorithm. It is used to generate the training data set for each of the base regressors in the proposed BOD prediction model.

C. *Principal Component Analysis(PCA)*

The training dataset for each base regressor in the proposed BOD prediction model is generated by rotated feature axis using PCA. As seen in research literature [25] [26] [27] PCA has been traditionally used as a dimensionality reduction technique. A set of correlated variables is transformed into a new set of uncorrelated variables. It captures the essence of the input features and creates a smaller feature set. The original features are then projected to this smaller set. Rotation forest uses PCA for creating rotated feature axes called principal components. These new axes are orthogonal and represent the directions with maximum variability in the dataset. These principal components are sorted in descending order of their significance, with the first component capturing the most variance in the data. The size of the data is then reduced by eliminating the weaker components and reconstructing an approximation of the original data using the stronger components. Rotation forest retains all the principal components generated by PCA, unlike dimensionality reduction. This ensures all discriminatory information about the data is retained and hence helps to improve the predictive accuracy.

D. Performance Metrics

In this paper predictive model performance is evaluated with two metrics first one is root mean square error (RMSE) and the second one is correlation coefficient. Correlation coefficient between actual BOD value and predicted BOD value calculated using equation 5 given below.

$$\text{Correlation coefficient} = \frac{S_{pa}}{\sqrt{S_p S_a}} \quad (5)$$

Where S_a and S_p are actual and predicted values respectively.

$$S_{pa} = \sum_i^n (p_i - \bar{p})(a_i - \bar{a}) \Big/ (N-1)$$

where $\bar{p}, \bar{a}$are the averages, respectively, and

$$S_p = \sum_i^N (p_i - \bar{p})^2 \Big/ (n-1) \quad \text{and}$$

$$S_a = \sum_i^n (a_i - \bar{a})^2 \Big/ (n-1) \quad (6)$$

If the values are perfectly correlated then the correlation coefficient is 1. A correlation coefficient of 0 indicates no correlation exists between them.

If p_i is the predicted value for i^{th} instance, a_i is the actual value for i^{th} instance and n is the total number of instances in the given data set, the root mean square error, RMSE is given by,

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(p_i - a_i)^2}{n}} \quad (7)$$

A model with a lower RMSE has a high predictive accuracy. The unit of RMSE is the same as the quantity being estimated namely BOD (mg/L).

IV. RESULTS AND DISCUSSION

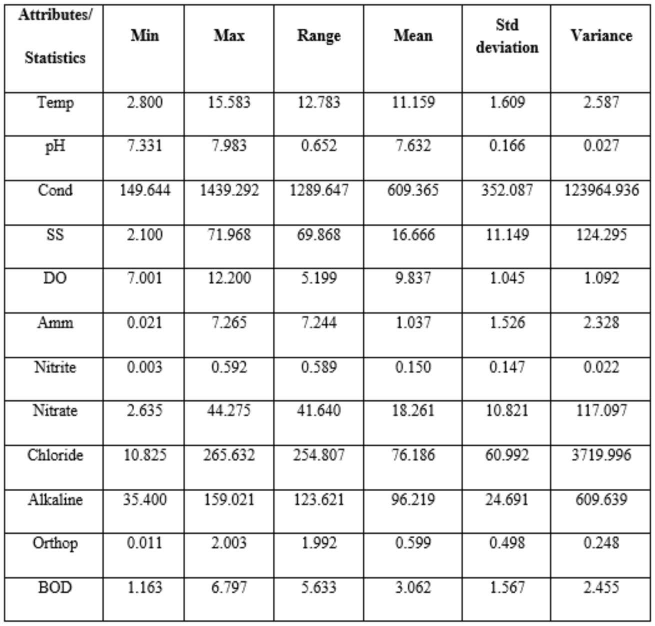
The data set was taken from the website of Department of Environment, Food and Rural Affairs, UK Government (DEFRA 2011) [29]. The data set contains average concentrations of various parameters of river water. In this paper a large annual data from 1980-2011 was considered from North-east region. The following parameters were collected from 128 water samples. Parameters include temperature (^{O}C), conductivity (µS/cm), pH, DO (mg/L), suspended solids (mg/L), nitrate (mg/L), ammoniacal nitrogen (mg/L nitrite (mg/L), chloride (mg/L),orthophosphate (mg/L), total alkalinity (mg/L) and BOD (mg/L). These attributes can be measured with the aid of sample in the case of DO, temperature and conductivity; for suspended solids gravimetry method was used and for other parameters standard titration techniques using common chemicals used for calculating the value. Above parameters were chosen because of their effects on measurement of BOD .There is a direct correlation between BOD and DO because BOD is measured by difference in levels of Dissolved Oxygen over the 5 day period. The amounts of nitrogen based compounds have shown their effects on BOD measurements. The temperature has an influence in the solubility of dissolved gases in water whereas pH gives an indicator of substances being alkaline or acidic. The conductivity measures the amount of ions present that affect microbial growth thereby indirectly influencing the BOD values.

A. Statistical Data Analysis

The descriptive statistics of the parameters used for water quality modeling in this paper is highlighted in Table 1.

Using10-fold cross validation in WEKA [28] a data mining tool a rotation forest ensemble for BOD prediction with model as base regressor is modelled. Cross validation is the most reliable test option for data driven models. In this technique, the data set is divided into 10 disjoint subsets and the model is built in 10 iterations. In each iteration, the model is trained using 9 subsets and tested using the left one subset. The predictive accuracy of the model in each iteration is then aggregated. The data set had missing values for some of the attributes especially alkalinity. As a pre-processing step these missing values were replaced by the mean of the other values of the attribute. The performance of this model is also compared with MLP neural networks -- a commonly used method for numeric prediction. The MLP neural network model in WEKA normalizes the attributes including the output attribute in order to improve the performance of the model.

Table 1. Descriptive Statistics of the Data Set



Attributes/ Statistics	Min	Max	Range	Mean	Std deviation	Variance
Temp	2.800	15.583	12.783	11.159	1.609	2.587
pH	7.331	7.983	0.652	7.632	0.166	0.027
Cond	149.644	1439.292	1289.647	609.365	352.087	123964.936
SS	2.100	71.968	69.868	16.666	11.149	124.295
DO	7.001	12.200	5.199	9.837	1.045	1.092
Amm	0.021	7.265	7.244	1.037	1.526	2.328
Nitrite	0.003	0.592	0.589	0.150	0.147	0.022
Nitrate	2.635	44.275	41.640	18.261	10.821	117.097
Chloride	10.825	265.632	254.807	76.186	60.992	3719.996
Alkaline	35.400	159.021	123.621	96.219	24.691	609.639
Orthop	0.011	2.003	1.992	0.599	0.498	0.248
BOD	1.163	6.797	5.633	3.062	1.567	2.455

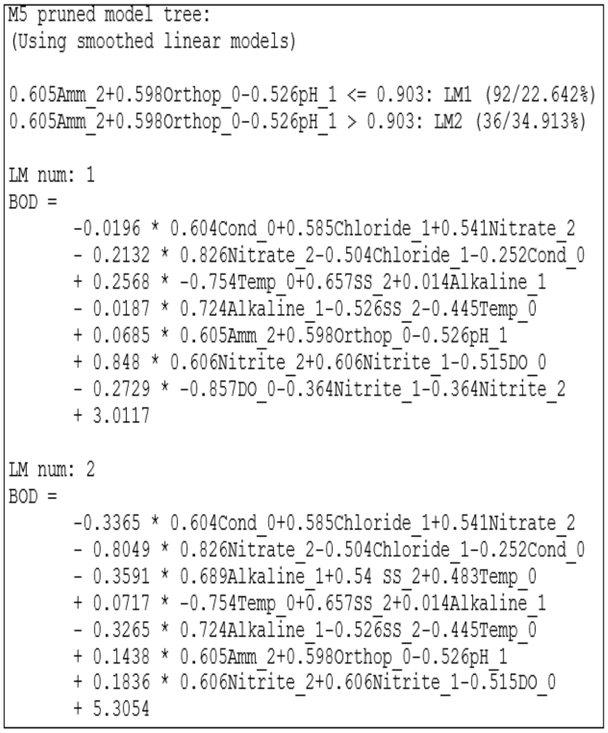
Table 2. Comparison of Predictive Accuracy

Model	Correlation Coefficient	RMSE
Rotation Forest with Model Trees as base regressors	0.9386	0.5388
Neural Networks	0.9168	0.6728

The parameter settings of this model are: training Time = 500, learning Rate = 0.3, validation Set Size = 0, seed = 0, validation Threshold = 20, momentum = 0.2,and hidden Layers = a, number of attributes. The Rotation forest ensemble is modeled using PCA as the projection filter. The percentage of instances eliminated from the training set while building the rotation matrix for each base regressor is 50%. Table 2 shows the performance comparison of the two models in terms of correlation coefficient and RMSE. The results in Table 2 show that the rotation forest ensemble with model trees as base regressors perform better than neural networks. The number of regression models formed at the leaf nodes of the decision trees built in each fold is shown in Table 3. The maximum number of decision paths in the resulting model is 9 and the minimum is two. Each decision path ends in a regression model that fits the subset of examples that reach the corresponding leaf node. The Model Tree built in fold 6 is illustrated in Figure 3

Table 3. Number of Regression Models for each Model Tree

Fold	1	2	3	4	5	6	7	8	9	10
No. of Regression Models	9	2	4	9	4	2	4	2	6	5



```
M5 pruned model tree:
(Using smoothed linear models)

0.605Amm_2+0.598Orthop_0-0.526pH_1 <= 0.903: LM1 (92/22.642%)
0.605Amm_2+0.598Orthop_0-0.526pH_1 > 0.903: LM2 (36/34.913%)

LM num: 1
BOD =
      -0.0196 * 0.604Cond_0+0.585Chloride_1+0.541Nitrate_2
      - 0.2132 * 0.826Nitrate_2-0.504Chloride_1-0.252Cond_0
      + 0.2568 * -0.754Temp_0+0.657SS_2+0.014Alkaline_1
      - 0.0187 * 0.724Alkaline_1-0.526SS_2-0.445Temp_0
      + 0.0685 * 0.605Amm_2+0.598Orthop_0-0.526pH_1
      + 0.848 * 0.606Nitrite_2+0.606Nitrite_1-0.515DO_0
      - 0.2729 * -0.857DO_0-0.364Nitrite_1-0.364Nitrite_2
      + 3.0117

LM num: 2
BOD =
      -0.3365 * 0.604Cond_0+0.585Chloride_1+0.541Nitrate_2
      - 0.8049 * 0.826Nitrate_2-0.504Chloride_1-0.252Cond_0
      - 0.3591 * 0.689Alkaline_1+0.54 SS_2+0.483Temp_0
      + 0.0717 * -0.754Temp_0+0.657SS_2+0.014Alkaline_1
      - 0.3265 * 0.724Alkaline_1-0.526SS_2-0.445Temp_0
      + 0.1438 * 0.605Amm_2+0.598Orthop_0-0.526pH_1
      + 0.1836 * 0.606Nitrite_2+0.606Nitrite_1-0.515DO_0
      + 5.3054
```

Figure 3. The Pruned Model Tree for Predicting BOD

The tree induced is a one-level decision tree. The two branches are the test conditions based on the chemical parameter ammoniacal nitrogen (Amm). The regression models corresponding to the two leaf nodes are LM 1 and LM 2 respectively. For each water sample the difference between the actual BOD and its BOD predicted by the Rotation Forest model is shown in Figure 4. Figure 5 shows the difference between the actual BOD and BOD for each water sample predicted by the MLP model.

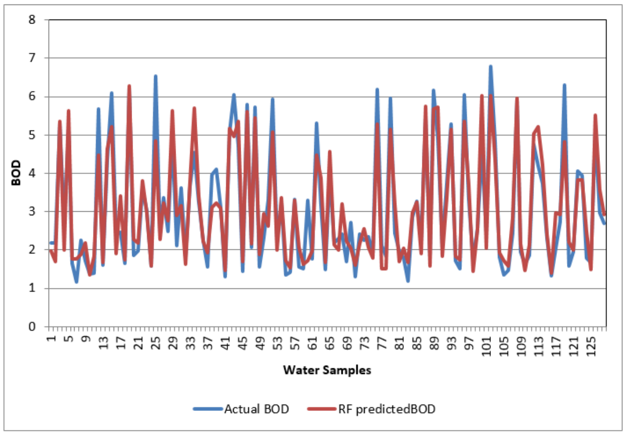


Figure 4 Comparison between the Actual BOD and BOD Predicted by Rotation Forest

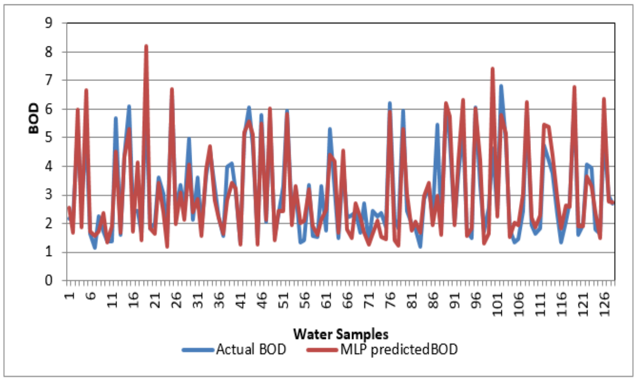


Figure 5 Comparison between the Actual BOD and BOD Predicted by MLP

It can be observed from Figure 4 and Figure 5 that the correlations between the actual BOD and BOD predicted by the rotation forest RF ensemble is more than the predictions made by MLP. The rotated feature axes using PCA for all the base regressors of Rotation Forest have significantly improved the diversity among all the base learners.

V. CONCLUSION AND FUTURE SCOPE

A rotation forest ensemble using model trees as base learners for predicting BOD in river water is proposed in this paper. The training dataset for the base regressors are constructed using rotated feature axes by PCA. The proposed model is also compared with the traditional neural network prediction model MLP. Experimental analysis on the available datasets has shown that the rotation forest ensemble has a high correlation coefficient and a low RMSE.

REFERENCES

[1] Clair N. Sawyer., Perry L., & McCarthy. 2003, *Chemistry for environmental Engineering and Science*, Tata McGraw Hill.

[2] Metcalf and Eddy. 2003.*Wastewater Engineering-Treatment and Reuse*, McGraw Hill. 4th edition.

[3] Tan P. Steinbach M. Kumar V. 2006. *Introduction to Data Mining*, Pearson Education.

[4] Musavi-Jahromi, S.H., and Golabi. M., 2008. Application of Artificial Neural Networks in the River Water Quality Modeling: Karoon River, Iran . *Journal of Applied Sciences*, 8: 2324-2328.

[5] Chadaphim P, Weeris T, Nagul C and Rajalida, 2016, Biochemical Oxygen Demand Prediction for Chaophrayariver using α-trimmed ARIMA model. *13th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE.*

[6] Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling and Software.* 15, 101–124

[7] [Masrur Ahmed A. A, 2017, Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs), *Journal of King Saud University – Engineering Sciences*,29 (2), 151-158.

[8] Masrur Ahmed, Syed Mustakim Ali Shah. 2017. Application of adaptive neuro-fuzzy inference system(ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River. *Journal of King Saud University – Engineering Sciences*,29(3), 237-243

[9] Palani, S. Shie-Yui Liong, Pavel Tkalich. 2008. An ANN application for water quality forecasting, *Marine Pollution Bulletin.* 56, 1586–1597.

[10] Dutta P and Chaki. 2012. A Survey of Data Mining Applications in Water Quality Management, *CUBE Intl. Information Technology Conference*, 470 -475.

[11] Chan, Kwonk-Wing, and Nitin Muttil. 2007. Data Mining and Multivariate statistical analysis for ecological system in coastal waters. Journal of Hydroinformatics. 9(4).

[12] Brydon, D.A., Frodsham, D.A. 2001. A model-based approach to predicting BOD[sub 5] in settled sewage. *Water Science & Technology*. 44 Issue 2/3, 9-15

[13] Rene, E R. and Saidutta, M. B. 2008. Prediction of Water Quality Indices by Regression Analysis and Artificial Neural Network.,*Int. J. of Environmental Research*, 2(2).183-188.

[14] Dominguez-Granda, L., Lock , K., and P. L. M. Goethals. 2011. Application of classification trees to determine biological and chemical indicators for river assessment: case study in the Chaguana watershed (Ecuador). *Journal of Hydroinformatics*,. 13(3). 489-499.

[15] Rodriguez. J. J., Ludmila I. Kuncheva, Carlos J. Alonso. 2006. Rotation Forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 28(10).1619-1630

[16] Kunchieva L.I and Rodriguez J.2007. An Experimental Study of Rotation Forest Ensembles, *LNCS, Springer-Verlag*,.459 – 468.

[17] Kotsiantis S.B, and Pintelas P.E. 2009. Local Rotation Forest of Decision Stumps for Regression Problems. *In 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT*.170-174.

[18] Lasota T, Luczak T and Trawinski B.2012. Investigation of Rotation Forest Method Applied to Property Price Prediction. *Artificial Intelligence and Soft Computing LCNS*, Springer-Verlag, 7267, 403-411.

[19] Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling and Software.* 15, 101–124

[20] Palani, S. Shie-Yui Liong, Pavel Tkalich. 2008. An ANN application for water quality forecasting, *Marine Pollution Bulletin.* 56, 1586–1597.

[21] Soman K.P and Diwakar S. 2006. *Insight into Data Mining: Theory and Practise*, PHI.

[22] Roiger, R.J., &Geatz, M.W., 2003. *Data Mining A Tutorial Based Primer*. Addison Wesley.

[23] Quinlan J. R. 1992. Learning with Continuous Classes, *Proceedings of 5th Australian Joint Conference on Artificial Intelligence,* World Scientific, Singapore, 343 – 348.

[24] Han J., and Kamber. M., 2001. *Data Mining: Concepts and Techniques*. Morgan Kaufmann

[25] Watcharapinchai N., Aramvith, S., Siddhichai, S., &Marukatat.S., 2008. Dimensionality Reduction of SIFT using PCA for Object Categorization. *2008 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2008)*, Thailand, 1-4.

[26] Lakhina, S., Joseph, S., and Verma. B., 2010. Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD. *Int. J. of Engineering Science and Technology,* 2(6),1790-1799

[27] Witten, I. H.and Eibe Frank. 2000. *Data Mining-Practical Machine learning tools and technology with Java implementations*, Morgan Kauffman.

[28] Hall. M., Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. 2009. The WEKA Data Mining Software: An Update, SIGKDD Explorations (2009), 11(1).

[29] Department of Environment, Food and Rural Affairs (DEFRA), 2011. UK Government website- http://data.gov.uk/dataset/river-water-quality-regions.

Authors Profile

Dr. J Alamelu Mangai pursed Bachelor of Engineering from Bharathidasan University in Computer Science and Engineering, Master of Engineering from Annamalai University in Computer Science and Phd from BITS-Pilani She is currently working as Associate Professor in Department of Computer Science and Engineering in Presidency University Bengaluru.She is a member of ISTE. she has published more than 7 research papers in reputed international journals and 7 conference paper it's also available online. Her main research work focuses on Data Mining-Application and Algorithms,Machine Learning. Shee has nearly 12 years of teaching experience and Research Experience.

Dr. Bharat B Gulyani pursed both Bachelor of Engineering in Chemical Engineering and Master of Engineering in Process Engineering and Plant Design from University of Roorkee.He has completed his Phd in the field of Chemical Engineering from University of Roorkee(now IIT Roorkee) in the year 1999.He is currently working as Associate Professor in the department of Chemical Engineering in BITS-Pilani Dubai Campus,Dubai.He is a member of IchemE,U.K and AIChE,USA.He has published more than 14 research papers in reputed journals and also more than 26 conference papers.His main research work focuses on Water Quality Management,Process Integration,Heat Exchanger Design,Energy Studies and Business Process Modeling.

Ms. Rashda Khanam pursed Bachelor of Technology from Biju Patnaik University of Technology in Computer Science and Engineering and Master of Technology from IIT(ISM) Dhanbad in Computer Science and Engineering.She is currently working as Assistant Professor in Department of Computer Science and Engineering in Presidency University Bengaluru since 2018.She has published more than 4 conference paper in national and International Conferences. Her main research work focuses on Cloud Computing,Optimization Approach,Machine Learning and Remote Sensing. She has nearly 10 month of teaching experience and 2 years of research experience.