The International Journal of Engineering and Science (IJES)
|| Volume || 7 || Issue || 9 Ver.I || Pages || PP 59-68 || 2018 ||
ISSN (e): 2319 – 1813 ISSN (p): 23-19 – 1805

# Anomaly Detection of River Data For Disaster Prevention

Md Saifur Rahman, Fahmida Pervin Brishty
*Department of Computer Science, Bangladesh University of Engineering and Technology*
*Institute of Information & Communication Technology, Bangladesh University of Engineering and Technology*
*Corresponding Author: Md Saifur Rahman*

**----------*ABSTRACT*----------**
*Anomaly in other word outlier detection is an invaluable technique to assess anypattern that is unconventional or unexpected with respect to other observation or event that. This approach is now being widely used in the detection of undesirable events such as network intrusions, bank fraud, and medical problems, errors in text or data and natural calamities. This paper particularly describes some anomaly detection algorithm to detect anomalous behavior in historical river water data collected from an Australian Database. The main feature we have concentrated on is the level of water level measured on daily basis from 1974 to till now. Maximum observations exhibit normal responses but there are also be very few observations of level acting anomalously within this dataset. There are nearly500 anomalies in final 15905 observations which are found from the simulation of different anomaly detection models. To reveal efficacy of various anomaly detection methods, Multivariate Outlier Detectionhas been implemented and compared to Statistical Parametric and Non-parametric detection,Cluster based Kmeans,Classification basedNeural Network and Support Vector Machine approach, Distance Based KNN approach and LOF density based method. Finally, we have presented a discussionon our findings and suggested actionable decision.*

***KEYWORDS-****KNN, Outlier, Parametric, Kmeans, SVM,Neural, Multivariate, Distribution, Likelihood, Q-Q, Box.*

---

Date of Submission: 25-08-2018 Date of acceptance: 08-09-2018

---

## I. INTRODUCTION

According to Hawkins, an observation turn out to be an outlier when differs so much from the other observations and suspicions are stimulates that it was produced in a different way. Some usual examples are unusual credit card purchase, sports, weather, commercial, financialevents etc. Outliers are different from the noise data in the way that noise is random error or variance in a measured variable. That's why before applying outlier detection, noise should be removed. The study of outlier detection is very interesting and it is now commonly being used in credit card fraud detection, telecom fraud detection, customer segmentation, medical analysis and many other sectors. Here we will apply different outlier detection techniques for detecting river water level anomalies over a period of 54 years. This data can give an overview about after what period water level can go beyond or below usual limit. Therefore is a useful interpretation for hydroelectric power generation, flood or draught forecasting, fisheries and other marine creatures supply and demand forecasting, environment monitoring and controlling, riverine transportation system projecting and so on. Outlier can be categorized into three major kinds: global, contextual and collective outliers.If an entityexpressivelymoves away from the rest then it is a situation of global or point anomaly and we need to find an appropriate measurement of deviation for it. An object is said to be conditionallyin other word contextually outlyingif it diverges meaningfully based on a selected context. The data we are attending falls into this type of anomaly with contextual and behavioral attributes. Another type- collective outlier arises when asubcategory of entity collectivelydifferssconsiderably from the whole entity, even if the individuals may not be outlier.


**Fig1: Global and Collective Outlier**

## II. MATERIALS AND METHODS

This paper considers a standard river water report dataset available in Australian Government Riverine database [7]. The data consists of historical water level, conductivity, discharge of Swan Hill Murray River from November, 1974 to June, 2018. The data is recorded on daily basis at 9:00 am. This information is used later for forecasting the new water level response. Variables having high percentage of missing values, Zero and NearZero-Variance, highly correlated variables were removed.For applying algorithm, the required transformation and coding have been carried out with R (version 3.5.1)[8]. Programming environment has been used along with RStudio 1.0.143 under the Windows 7 Operating System with RAM of 4 GB and Intel Core i5 processor. The implementation of the prediction algorithm has been done with the help of an MATLAB 2018. Two ways to categorize outlier detection methods: Based on whether user-labeled examples of outliers can be obtained: Supervised, semi-supervised vs. unsupervised methods[18].

There are mainly two ways to classify outlier detection procedure. If the outliers are user-labeled, then supervised, semi-supervised and unsupervised methods are applicable. Provided that the analyst have notions about normal data and outliers then statistical, proximity-based, and clustering-based methods are practical. For supervised outlier detection we treat the situation as a classification problem. Samples examined by domain experts used for training & testing. Then normal objects are modelled& those not matching the model arereported as outliers, or outliers are modelled and those not matching the model are treated as normal.The challenges are imbalanced classes, boosting the outlier class and making up some artificial outliers. For modelling we need to catch as many outliers as possible andassume the normal objects are somewhat ``clustered'' into multiple groups, each having some distinct features and outlier is expected to be far away from any groups of normal objects although we cannot detect collective outlier effectively. Even if we don't observe strong patterns in normal objects, the collective outliers may exhibit high similarity in a small area. Unsupervised methods fails to detect many real outliers. Supervised methods can be more effective. Many clustering methods can be adapted for unsupervised methods where first clusters are built, then outliers which do not belonging to any cluster. But there are few problems such as in clustering it is hard to distinguish noise from outliers and it is also costly since first clustering as for far less outliers than normal objects. Some newer methods which tackle outliers directly are semi-supervised outlier detection. It can be regarded important if some labeled normal objects are available and if we can use the labeled examples and the proximate unlabeled objects to train a model for normal objects. Those not fitting the model of normal objects are detected as outliers. Outlier detection improvement is very feasible with the help from models for normal objects learned from unsupervised methods.

## III. RESULTS

Our work consists of the following major steps: data visualization, statistical method, proximity based method, clustering based method [18].It shows variation in trend of water level in between 0.5 to 4.8 from 1974 to 2018. Original data is plotted like this.

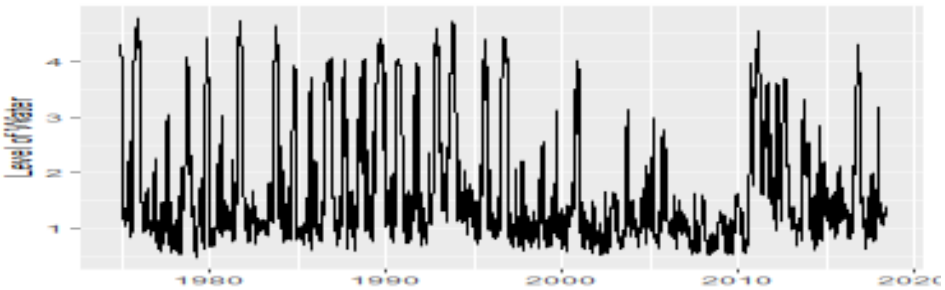


**Fig.2.Historical water level data over a period of 45 years.**

### A. STATISTCAL APPROACH

Statistical methods or model-based methods assume that the normal data follow some statistical stochastic model. The data not following the model are termed as outliers. First a particular distribution like Gaussian, Binomial or Exponential etc. is used to fit the normal data. For each object y in region R, we estimate the probability of y fits the specific distribution. If the probability is very low, y is unlikely generated by the particular model, thus it is an outlier. Efficiency of statistical outlier detection depends on whether the assumption consists of real data. This assumes that the objects in a data set are caused by a stochastic process and are divided into two categories: parametric vs. non-parametric.

**B.1.PARAMETRIC APPROACH I: DISTRIBUTION BASED**

Parametric method assumes that the normal data is generated by a parametric distribution with parameter θ. The probability density function of the parametric distribution $f(x, \theta)$ gives the probability that object $x$ is generated by the distribution. Smaller probability denotes that the data is more likely to be an outlier. Non-parametric method does not assume a-priori statistical model and determine the model from the input data. Process is not parameter free in a sense that it takes into account some number and parameters. But this is flexible and not fixed like histogram and kernel density estimation. Parametric based univariate outlier detection has been done on Normal Distribution. The river water data is univariate as it involves only one attribute water level. Assuming that data are generated from a normal distribution, we learn the parameters from the input data, and then identify the points with low probability as outliers. The maximum likelihood method is used to estimate μ and σ. Taking derivatives with respect to μ and $\sigma^2$, we derive the following maximum likelihood estimates [18].

$$\ln \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^{n} \ln f(x_i|(\mu, \sigma^2)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

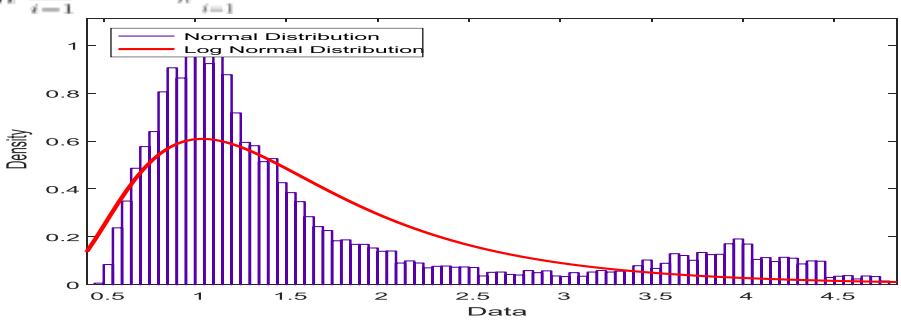


**Fig.3. Density of water level data over normal and log-normal distribution.**

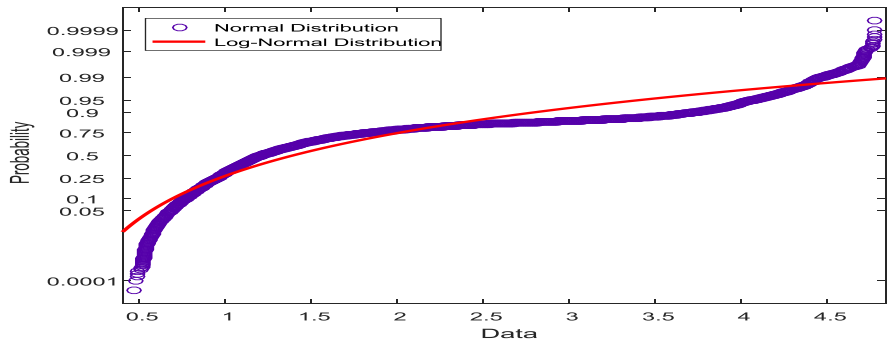


**Fig.4. Probability of water level data over normal and log-normal distribution.**

The x-axis, in the above plot, represents the water level data and the y-axis, probability density of the observed data. The density curve for the normal distribution is shaded in 'violet' and log normal distribution is shaded in 'red'. The probability density for the normal distribution is calculated from the practical data, while both normal and log-normal distribution is computed based on the detected mean and standard deviation of the water level data. Outliers could be acknowledged by calculating the occurrence probability of an observation or calculating how far the observation is from the mean. In the above case, if we assume a normal distribution, there could be how much outlier level especially for observations is beyond 4.5. The log-normal plot is much better than normal distribution as the underlying actual distribution has characteristics of a log-normal distribution. The parameters of the data are inferred by fitting a curve to the data though a change in the underlying parameters like mean and/or standard deviation due to new incoming data will change the location and shape of the curve. Any alteration in the parameters of a distribution evidently influences the identification of outliers.

**B.2.PARAMETRIC APPROACH II. THE GRUBB'S TEST BASED**

It is also a univariate outlier detection technique normal distribution. It is also called maximum normalized residual test. For each object x in a data set, we need to calculate its z-score. Then x is considered as an outlier if

$$z \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

where, $t^2_{\alpha/(2N),N-2}$ is the value taken by a t-distribution at a significance level of α/(2N), and N is the of objects in the data set. After running the algorithm in R-studio, we found G = 2.96, U = 0.99, p-value = 1 under alternative hypothesis. And any value higher than 4.777 is considered as an outlier here [18].

### B.3. PARAMETRIC APPROACH III. DETECTION OF MULTIVARIATE OUTLIERS

For Multivariate data two or more attributes or variables we transform the outlier detection task into a univariate outlier detection problem. First, we compute Mahalaobis distance. Let ō be the mean vector for a multivariate data set. Mahalaobis distance for any object o to ō is Maha Dist(o, ō) = $(o - \bar{o})^T C^{-1}(o - \bar{o})$ where C is the covariance matrix. Then we use the Grubb's test on this measure to detect outliers. Outlier detection techniques will normalize all of the data, so the mismatch in scaling is of no consequence. Close attention must still be called to the variables themselves. It is important to note that the first variable corresponds to an identification number rather than a data point and should not be included in outlier detection analysis. For outlier visualization of 2-dimensional data, color.plot function of is used where the inputs are data frame or matrix, amount of observations to be used for mcd (Minimum Co-variance Determinant) calculation and amount of observations used to calculate the adjusted quantile. The plot generated by this function represents the euclidean distance as a color, where blue represent small distances, and red represent large distances from the data set minimum. Mahalanobis distances are represented by their symbol. The function also draws four ellipsoids where the Mahalanobis Distances are constant. The constant values correspond to the 25%, 50%, and 75% and adjusted quantiles. What this technique lacks more than anything is scalability. The data that can be represented in this way is only limited to 2-dimensional. In the present, where big-data is becoming the norm, a multivariate solution is preferable [8].

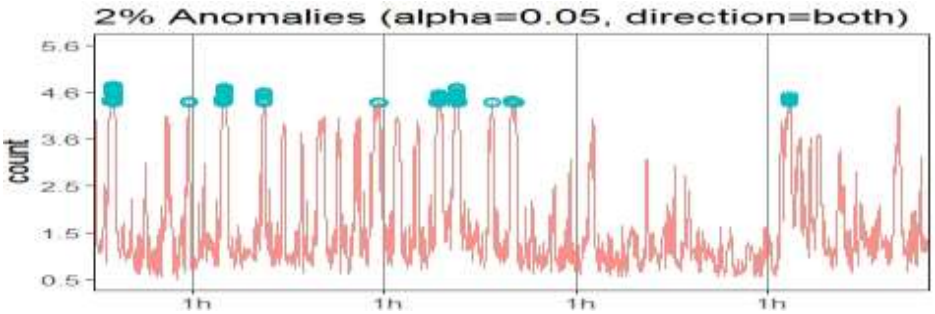


**Fig.5. River data anomalies detected by multivariate outlier detection method.**

The input time series plot exhibits that it has positive and negative anomalies case. Besides, local anomalies due to series' seasonality cannot be detected using the traditional approaches. The anomalies detected using the proposed technique is annotated on the plot.

### B.4. PARAMETRIC APPROACH III. $\chi^2$ –STATISTICS

In $\chi^2$ –statistic: $\boldsymbol{X^2} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

Where, $E_i$=the mean of the *i*-dimension among all objects, and n=the dimensionality. If $\chi^2$ –statistic is large, then object $o_i$ is an outlier for chi-squared test. After simulating the data in R-studio, we get X-squared = 1.2217, p-value = 0.269 with alternative hypothesis for lower boundary. And as a result value lower than 0.473 is marked as an outlier. Similarly, for upper boundary technique we have got X-squared = 8.7357, p-value = 0.00312 with alternative hypothesis and value higher than 4.777 is considered as an outlier.

### B.5. NON-PARAMETRIC APPROACH I. DETECTION USING HISTOGRAM

The histogram model of normal data can be easily learned from the input data without any prior knowledge regarding the data. The main advantage is that it often makes fewer assumptions about the data, and thus can be applicable in more scenarios. Histogram fitted with our particular dataset is shown below:

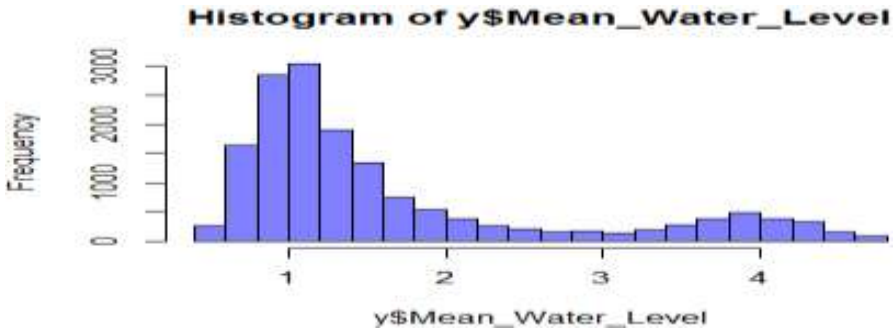


**Fig.6. Histogram for water data outlier detection.**

Figure shows the histogram of water level over a period of 45 years. Level in the amount of higher than 4.5 is an outlier, since only 0.2% observations have an amount higher than 4.5. But the problem is that it is hard to choose an appropriate bin size for histogram. If the bin is too small then normal objects in empty/rare bins can cause false positive. Or if the bin size is too big, outliers in some frequent bins can be treated as false negative. There is a solution to adopt kernel density estimation to estimate the probability density distribution of the data.If the estimated density function is high, the object is likely normal. Otherwise, it is likely an outlier [19].

**B.6. NON-PARAMETRIC APPROACH II. DETECTION USING BOX AND QQ PLOT**

If the mean accurately represents the center of the distribution and the data set is large enough, parametric approach could be used whereas if the median represents the center of the distribution, non-parametric approach to identify outliers is suitable. Dealing with outliers in a multivariate scenario becomes all the more tedious. A non-parametric method could be used to identify outliers in such a case. Box plot[19] shows the relationship between a numerical y-variable and a grouping x-variable by using the five number summary—minimum, first quartile (Q1), median, third quartile (Q3), maximum. IQR is interquartile range. Any point falling outside of LAV and UAV are marked as outliers. The tooltip label includes additional information about the outlier which is different compared to all other data points in the plot. In addition to the above, the equation provides lower adjacent value (LAV) and upper adjacent value (UAV) defined as follows:

LAV = Q1 – 1.5 * IQR, UAV = Q3 + 1.5 * IQR

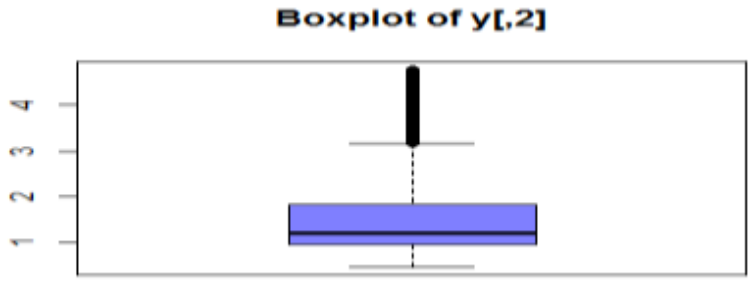


**Fig. 7. Box plot showing the outliers in the river water level data.**

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. The advantages of the q-q plot are that sample sizes do not need to be equal.

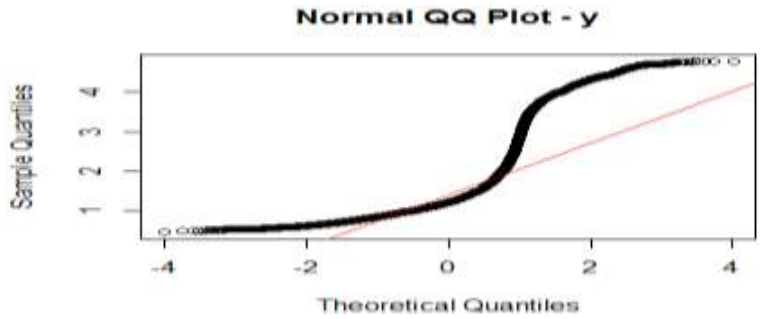


**Fig. 8. Q-Q plot denoting the presence of outlying data in river water data.**

The q-q plot [9] is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.This q-q plot shows that some data do not appear to have come from populations with a common distribution and some values are significantly higher than the corresponding normal values. The q-q plot is formed by vertical axis which is the estimated quantiles from data set and horizontal axis that is the theoretical quantiles from data set. Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample test.

**B. SUPERVISED KMEANS CLUSTERING BASED OUTILIER DETECTION**

We have used weka simple KMeans clustering method [17] for outlier selection. For clustering using the k means algorithm, we have used the Euclidean distance instead of the Manhattan distance. The random number seed has been used. Display standard deviations of numeric attributes and counts of nominal attributes are not done. The number of execution slot is chosen as 1. Minimum canopy density is chosen as 2 for initialization and/or speedup. This is the minimum T2-based density below which a canopy will be pruned during periodic pruning. Number of cluster is set as 2, number 1 is the correct cluster and number 2 is the misclassified cluster. Clusterer capabilities are not checked before clusterer is built. Max iterations has been set at 500 preserving order of instances. Canopy periodic pruning rate is set 10000 for initialization and/or speedup this is how often to prune low density canopies during training. Initialization method is chosen at random. The distance function to use for instances comparison here is Euclidean Distance. Within 11 numbers of iterations, within cluster sum of squared errors is 1469.057. Cluster centroids' index are chosen as 7953, 8288, 6522, 6493 and 579. Time taken to build model is 0.03 seconds. 81% of clustered Instances belong to normal group while 19% are outliers.

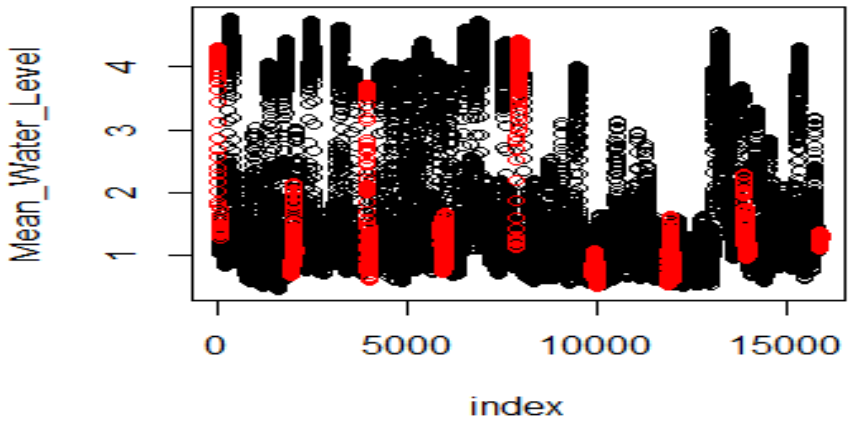


**Fig.9. Black color showing normal data and red color showing anomalous data in kmeans clustering.**

**B. UNSUPERVISED MULTILAYER PERCEPTRON OUTLIER DETECTION**

Multilayer perceptron classifier uses back propagation to classify instances. The network can also be monitored and modified during training time and the nodes in this network are all sigmoid. Here a seed resets the random number generator setting the initial weights of the connections between nodes, and also shuffles the training data. We used 0.2 momentums to the weights during updating. Nominal to binary filter will preprocess the instances with the filter to improve performance if there are nominal attributes in the data. 3 hidden layers were

defined for neural network. Validation threshold is set to 20 for terminating validation testing. The value is used to show how many times in a row the validation set error can get worse before the end of training. Alterations to the neural network can only be done while the network is not running. This also applies to the learning rate and other fields on the control panel. Normalize attributes is used for normalizing the attributes. This could help to improve performance of the network by normalizing nominal attributes. A batch size of 100 has been used as the preferred number of instances to process if batch prediction is being performed. More or fewer instancesmay be provided, but this gives implementations a chance to specify a preferred batch size. Decay causes the learning rate to decrease and divide the starting learning rate by the epoch number, to determine what the currentlearning rate should be. This progresses the performance and also halts the network from deviating from the goal output. Training time is set to 500. Learning rate is set to 0.3 for weights updating. Time taken to build model is 1.45 seconds and time taken to test model on training data: 0.05 seconds. After classifying on 15905 instances, it detected a large number of outlier shown in the graph.

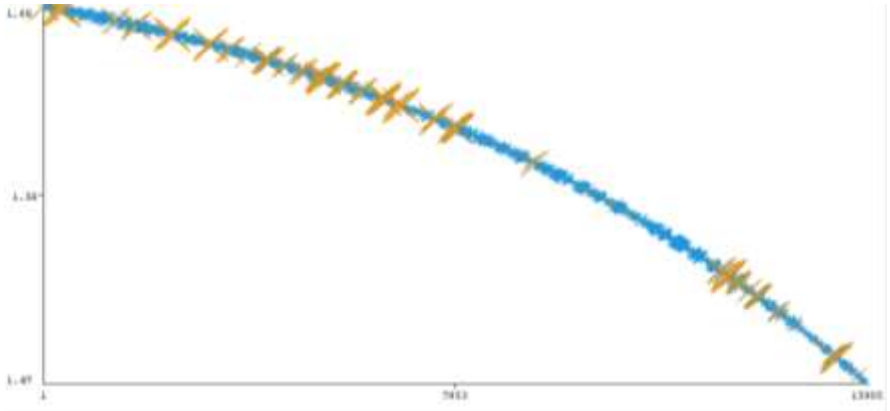

**Fig.10. Dot showing normal data and cross showing anomalous data in artificial neural networking based detection.**

### B. SUPERVISED TREE BASED OUTLIER DETECTION

For detecting outlier weka classifiers trees REP Tree has been used which is fast decision tree learner and uses information gain/variance along with pruning back fitting [16][17]. Missing values are dealt with by splitting the corresponding instances into pieces. The seed 1 is used for randomizing the data. The minimum total weight of the instances in a leaf is set as 2. Besides, 3 number folds determine the amount of data used for pruning. The preferred number of instances is selected as 100 to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size. Pruning is performed without spreading initial count across all values instead of using the count per value. Max depth is set at -1 for showing no restriction. The minimum proportion of the variance is set at 0.001 on all the data that needs to be present at a node in order for splitting to be performed in regression trees. With splitting 66.0% train, remainder test; size of the tree has become 3527 and it took 0.03 seconds to build the model. Finally we found 11.11 % relative absolute error in other words anomalies.

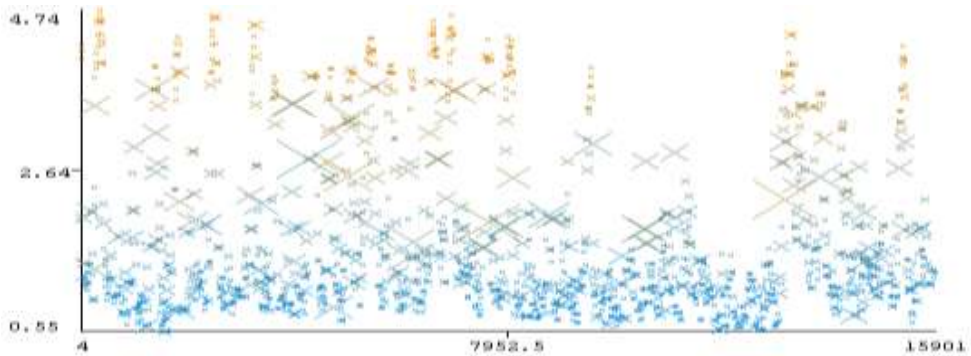


**Fig.11. Dot showing normal data and cross showing anomalous data in tree based outlier detection.**

### B. SUPERVISED KNN OUTLIER DETECTION

Weka lazy IBk classifier has been used for K-nearest neighbor classifier [16][17]. We select value of K as 10 based on cross-validation which is the number of neighbors. Batch size is the preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives

implementations a chance to specify a preferred batch size which is chosen as 100. No distance weighting method has been used. Nearest neighbor search algorithm is used as neighbor search. LinearNN Search. Window size is set at a value of 0 to signify no limit to the number of training instances. Mean squared error is not used for doing cross-validation for regression problems. It can address Binary class, Date class, Missing class values, Nominal class, Numeric class and attributes with updateable classifier, weighted instances handler. Among 15905 instances, 10-fold cross-validation was done on full training set classifier model using 10 nearest neighbor. 0.01 second was taken to build model and about 7.25 % relative absolute error or outlier was detected.

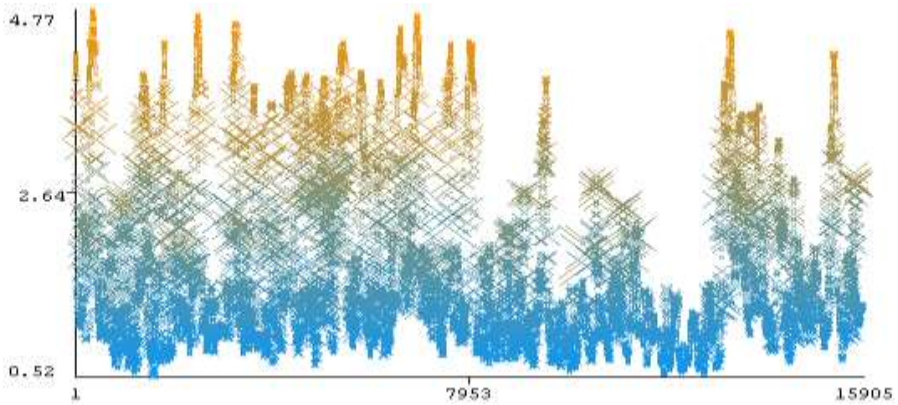


**Fig.12. Dot showing normal data and cross showing anomalous data in KNN based anomaly detection.**

**B. SUPERVISED SVM BASED ONE-CLASS OUTLIER DETECTION**

We have used weka classifiers SMO regression algorithm for implementing the support vector machine [21]. The algorithm has been selected by setting the Reg Optimizer. The parameters that we have used for tuning the model are the complexity parameter C. Two decimal places are used for the output of numbers in the model. The batch size, preferred number of instances to process is used as 100. We have used linear basis kernel and then debugged. We didn't use any filter for data transformation. This algorithm efficiently learns any type of date class, missing class values, numeric class. The benefit is that it can work with different types of attributes like binary attributes, empty nominal attributes, missing values, nominal attributes, numeric attributes, and unary attributes etc. After running the support vector algorithm one-classification and linear kernel with parameters gamma: 0.5, nu: 0.09; we finally got 721 number of support Vectors for classifying the outliers. In the below figure the red ones are the outlying data.

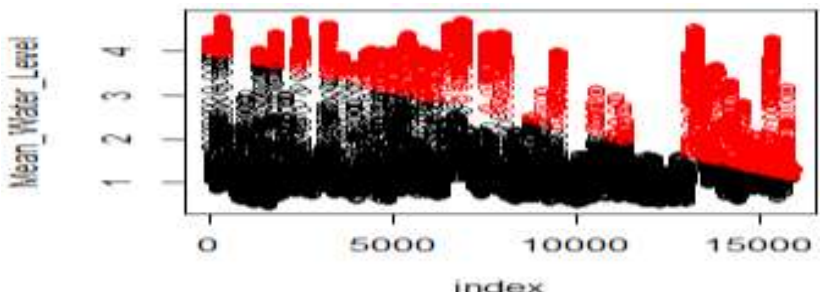


**Fig.13. Black color showing normal data and red color showing anomalous data in support vector machine.**

**B. LOF DENSITY BASED OUTLIER DETECTION**

The proximity based outlier detection techniques falls under two major categories namely the density-based and the distance-based outlier detection. Generally, in the proximity-based outlier detection technique an object is considered to be an outlier if it is distant from most other points.This approach is more general and simpler than the statistical approaches, since it is easier to determine a meaningful proximity measure for a data set than to determine its statistical distribution.In this lesson we will only focus on discussing about the density based outlier detection technique.A simplest way to measure whether an object is distant from most other points in the dataset is to use the distance to the k-nearest neighbor or the LOF (Local Outlier Factor).The LOF methodis based on scoring outliers on the basis of the density in the neighborhood. This technique is based on a parameter known as outlier score. The outlier score of an object is the reciprocal of the density in the object's neighborhood where density is the average distance to the k-nearest neighbors.In the LOF technique the local

density of a point is compared with that of its neighbors. If the former is significantly lower than the latter i.e. if LOF is greater than one, then the point is in a sparser region than its neighbors, which suggests it to be an outlier.The only limitation of LOF is that it works on numeric data only.The complexity of this algorithm is 0 (N2) and another challenge of this algorithm is to select a right value of k which is not very obvious.In the R, in package DMwR there is a function lofactor() which computes the local outlier factors using the LOF algorithm[20].

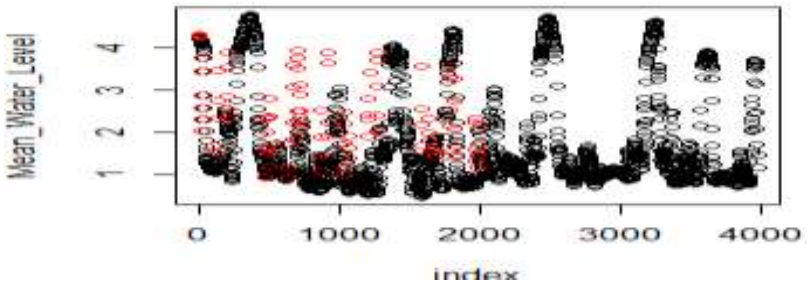


**Fig.14. Black color showing normal data and red color showing anomalous data in LOF based detection.**

The index reduced from 15900 to 4500 as we down sample the data as we don't have such compiling capacity to detect outlier. The quintile based on density is estimated as at 0% 0.9416643, at 25% 0.9999740, 50% 0.9999994, 75% 1.0000276, 100% 1.1505695.

## IV. DISCUSSIONS

We applied different algorithms for modeling normal objects and outliers properly. But all of them didn't show good performance in detecting outliers. It is hard to specify all possible normal behaviors in a single application. So, we need to apply several approaches on the same data to get a desired output level. The margin between normal and outlier objects is often a different colored area. Which type of distance measure is being used among objects and the model of relationship among objects are often algorithm-dependent. But the main thing we should do before outlier detection is that we must handle noise with care before. That's because noise may deform the normal objects and distort the distinction between normal objects and outliers. Besides, it may reduce the effectiveness of outlier detection by concealing outliers. Another thing is that we must gather understand ability regarding the justification of the detection. The degree of an outlier is specified as the unlikelihood of the object being generated by a normal mechanism.

## V. CONCLUSION

This paper introduces some specific routes for quantifying river data anomaly. The required steps are easier to implement and understandable, don't require standby technical personnel so in other word less expensive and more accurate. This application can help predicting hydro-electric power which ensures a cleaner environment with unpredictable power generation. Besides, the experiment can accurately forecast flood and river transportation system thus can help prevention. In near future, different types of time series reports from other databases may be used for further validation of the proposed method. With a predictable water-level based hydro-electric power generation, the fuel based generators can be adjusted with a pre-planned way.

## REFERENCES

[1]. Eskin, Eleazar. "Anomaly detection over noisy data using learned probability distributions." In In Proceedings of the International Conference on Machine Learning. 2000.
[2]. Stibor, Thomas, Jonathan Timmis, and Claudia Eckert. "A comparative study of real-valued negative selection to statistical anomaly detection techniques." In International Conference on Artificial Immune Systems, pp. 262-275. Springer, Berlin, Heidelberg, 2005.
[3]. Madhuri, G.S. and Rani, M.U., 2018. Anomaly Detection Techniques
[4]. Chandola, V., Mithal, V. and Kumar, V., 2008, December. Comparative evaluation of anomaly detection techniques for sequence data. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 743-748). IEEE.
[5]. Mell, P., Hu, V., Lippmann, R., Haines, J. and Zissman, M., 2003. An overview of issues in testing intrusion detection systems.
[6]. Lane, T.D., 2000. Machine learning techniques for the computer security domain of anomaly detection.
[7]. "RiverMurrayData,"[Online].Available:https://riverdata.mdba.gov.au/swan-hill
[8]. "RPubs - Anomaly Detection Tutorial,"[Online].Available:https://rpubs.com/Treegonaut/301942.
[9]. "Week 7 Project Anomaly Detection,"[Online].Available: https://rpubs.com/jneville001/272242.
[10]. "Fitting a Model by Maximum Likelihood,"[Online].Available:https://www.r-bloggers.com/fitting-a-model-by-maximum-likelihood/
[11]. "RProgramming/MaximumLikelihood,"[Online].Available:https://en.wikibooks.org/wiki/R_Programming/Maximum_Likelihod.
[12]. "anomalyDetectionpackage|RDocumentation,"[Online].Available:https://www.rdocumentation.org/packages/anomalyDetection/versions/1.0

[13]. "A tutorial on outlier detection techniques | R-bloggers,"[Online].Available: https://www.r-bloggers.com/a-tutorial-on-outlier-detection-techniques/
[14]. "Distance-based Outlier Detection in Data Streams,"[Online].Available: https://infolab.usc.edu/DocsDemos/p1184-tran.pdf
[15]. Stibor, T., Timmis, J. and Eckert, C., 2005, August. A comparative study of real-valued negative selection to statistical anomaly detection techniques. In International Conference on Artificial Immune Systems (pp. 262-275). Springer, Berlin, Heidelberg.
[16]. Patcha, A. and Park, J.M., 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer networks, 51(12), pp.3448-3470.
[17]. Hodge, Victoria, and Jim Austin. "A survey of outlier detection methodologies." Artificial intelligence review 22, no. 2 (2004): 85-126.
[18]. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
[19]. "Outlier detection with boxplot.stats function in R"[Online].Available::https://www.r-bloggers.com/outlier-detection-and-treatment-with-r/
[20]. "Outlier detection with Local Outlier Factor with R"[Online].Available::http://datatechnotes.blogspot.com/2017/12/outlier-detection-with-local-outlier.html
[21]. "Outlier check with SVM novelty detection in R,"[Online].Available::http://datatechnotes.blogspot.com/2018/01/outlier-check-with-svm-novelty.html

Md Saifur Rahman"Anomaly Detection of River Data For Disaster Prevention "The International Journal of Engineering and Science (IJES) 7.9 (2018): 59-68