ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Data Analysis in Forecasting Lakes Levels Using K-Medoid Clustering

Manali Shukla [1], Megha Seth [2]

M. Tech Student, Department of Computer Science, Rungta College of Engineering and Technology, Bhilai, Chattisgarh, India[1]

Reader, Department of Computer Science, Rungta College of Engineering and Technology, Bhilai, Chattisgarh, India[1]

ABSTRACT: Lakes are lively system those are insightful to confined climate and to land-use amendment in the neighboring site. Several lakes obtain their water primarily from rainfall, some are conquered by drainage overspill, and several others are illicit by land water systems. As time grows, the areal amount and profundity of water in lakes are indication of effect in climatic factors some of them are rainfall, emission, temperature, and airstream speed. Fluctuations of lake level diverge with the water equilibrium of the lake and their catchment, and might be, in assured cases, reflect changes in shallow groundwater resources. Prevalent surface fresh water system on the globe. The seasonal, monthly and yearly surface water level of the lakes alters in retort to an assortment of factors. In this paper we will use k-medoid clustering and multiple regression technique for lake level forecasting. By use of clustering data mining technique we will classify our data. As lake level changes as per time due to some weather conditions if we efficiently classify historic data. Further we will apply Cubic SVM classifier over the predicted data to compute consistency of data predicted.

KEYWORDS: Regression; Hadoop; K-Means; ARIMA.

I. INTRODUCTION

Forecasting is the process of making predictions or estimating of the future based on past and present data. Forecasting provides information about the future events and acts like a planning tool for the organization. Forecasting is the procedure of making certain assumptions based on the management's knowledge, experience, and judgment. Number of statistical techniques used by forecasting that's why we also called it as statistical analysis. Importance of forecasting involves following points:-

- Forecasting provides relevant and reliable information about the past and present events and the likely future events. This is necessary for sound planning.
- It gives confidence to the managers for making important decisions.
- It keeps managers active and alert to face the challenges of future events and the changes in the environment. [20]

Limitations of forecasting involves following points:-

- The collection and analysis of data about the past, present and future involves a lot of time and money. Therefore, managers have to balance the cost of forecasting with its benefits. Many small firms don't do forecasting because of the high cost.
- Forecasting can only estimate the future events. It cannot guarantee that these events will take place in the future. Long-term forecasts will be less accurate as compared to short-term forecast.
- Forecasting is based on certain assumptions. If these assumptions are wrong, the forecasting will be wrong. Forecasting is based on past events. However, history may not repeat itself at all times.
- Forecasting requires proper judgement and skills on the part of managers. Forecasts may go wrong due to bad judgement and skills on the part of some of the managers. Therefore, forecasts are subject to human error. [20]

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

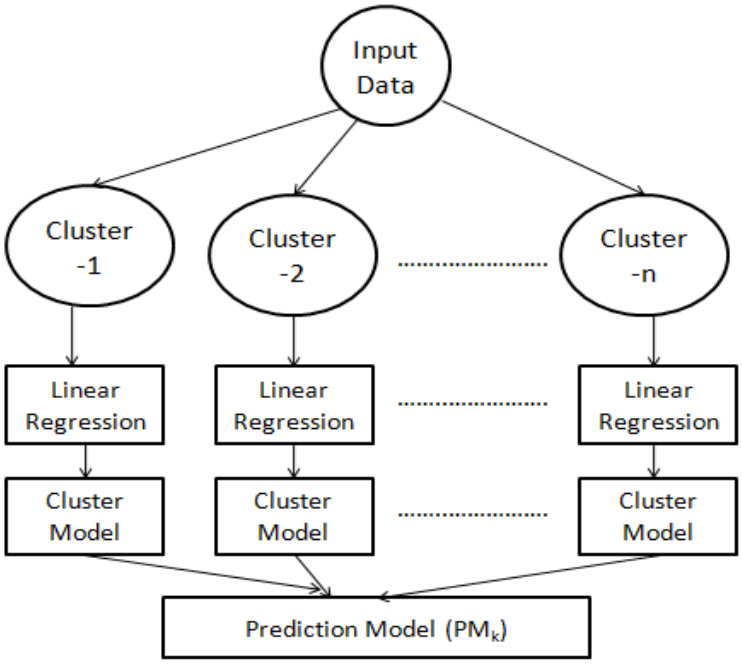

Fig.-1 Prediction Model

II. MOTIVATION

The ease of use and convenience of near real-time water level data at manifold locations for every of the lakes was a key constituent for the accomplishment of the prepared forecasting system. The dataset of chronological water levels and wave heights were worn to standardize the hydrodynamic and wave models for apiece lake, while the near real-time water level data was used for data absorption on the hydrodynamic models.

Correct prediction of water level vacillation is significant in lake management due to its noteworthy impacts in an assortment of aspects. This attempt to examine and suggest a computer platform which can be used in enduring for reservoirs level statement comes very useful. It's recognized that the system is with no trouble scalable to include supplementary knowledge for different reservoirs and always perform professionally.

Forecasting of water lake level can be useful for:-

- Assessing the severity of drought
- Predicting where/when there is a risk of flooding
- Limiting the impact of flooding and drought by enabling government and other agencies to put emergency response plans into operation
- Identifying sensitive and risk prone areas for flooding and drought throughout the province[4]

III. LITERATURE SURVEY

Ozgur Kisi et. al. [2] discussed that Forecasting lake level at various horizons is a critical issue in navigation, water resource planning and catchment management. In this article, multistep ahead predictive models of predicting daily lake levels for three prediction horizons were created. The models were developed using a novel method based on support vector machine (SVM) coupled with firefly algorithm (FA). The FA was applied to estimate the optimal SVM parameters. Daily water-level data from Urmia Lake in northwestern Iran were used to train, test and validate the used

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

technique. The prediction results of the SVM–FA models were compared to the genetic programming (GP) and artificial neural networks (ANNs) models. The experimental results showed that an improvement in the predictive accuracy and capability of generalization can be achieved by the SVM–FA approach in comparison to the GP and ANN in 1 day ahead lake level forecast. Moreover, the findings indicated that the developed SVM–FA models can be used with confidence for further work on formulating a novel model of predictive strategy for lake level prediction [Ozgur Kisi 2015].

Hakan Tongal et. al. [18] Told that lake water level forecasting is very important for an accurate and reliable management of local and regional water resources. In this study two nonlinear approaches, namely phase-space reconstruction and self exciting threshold autoregressive model (SETAR) were compared for lake water level forecasting. The modeling approaches were applied to high-quality lake water level time series of the three largest lakes in Sweden; Vanern, Vattern, and Malaren. Phase-space reconstruction was applied by the k-nearest neighbour (k-NN) model. The k-NN model parameters were determined using autocorrelation, mutual information functions, and correlation integral. Jointly, these methods indicated chaotic behaviour for all lake water levels.

Anne H. Clites et. al. [12] discussed about why Monitoring Water Levels required. They told the recent surge in water levels has provided relief to systems and economic sectors stressed by hydrologic extremes. The prolonged period of low water conditions preceding the recent surge, for example, catalyzed demands for new structures designed to reduce flow rates through the St. Clair River and increase water levels on Lake Michigan-Huronthe recent surge has changed the context of the debate over the benefits and the urgency of putting these structures in place. Internationally coordinated seasonal water level forecasts through the summer of 2015 indicate that monthly average water levels are likely to follow their typical seasonal trends at above-average levels. Beyond that time frame, however, drivers of regional climate variability that can significantly impact regional water budgets and lake water levels remain difficult to predict. The recent rise in water levels on Earth's two largest freshwater surfaces and the preceding period of below-average levels forced us to apply data prediction technique over Lake water level forecasting. [Anne H. Clites 2013].

S.No.	Author/Year	Method Used	Description
1.	Prashant Shrivastava et. al. Cloud Based Software Platform for Big Data Analytics In Water Reservoir Level Forecasting IOSR-JCE 2014 [14]	Autoregressive Integrated Moving Average (ARIMA) algorithmic	It stores the historic huge data set inside massive information storage and uses big data technologies to review behaviour and predict the future levels by applying data-driven analytics and data mining concepts.
2.	Chih-Chieh Young et. al. Predicting the Water Level Fluctuation in an Alpine Lake Using Physically Based, Artificial Neural Network, and Time Series Forecasting Models Hindawi Publishing Corporation 2015 [3]	Artificial neural network (ANN) model (back propagation neural network, BPNN) & ARMAX	Four modelling approaches (the three-dimensional hydrodynamic model, ANN model, ARMAX model, and the combination model) have been implemented to predict water level fluctuation
3.	Hakan Tongal et. al. Phase-space reconstruction and self-exciting threshold modeling approach to forecast lake water levels Springer-Verlag Berlin Heidelberg 2013[18]	k-nearest neighbour (k-NN) model & SETAR model	A comparison of two nonlinear model approaches was made. Author used the k-NN approach and SETAR model for prediction of water levels for the three largest lakes in Sweden.
4.	Shubhendu Trivedi e. al. The Utility of Clustering in Prediction Tasks Centre for Mathematics and Cognition gran 2011 [16]	k-means	Observed that use of a predictor in conjunction with clustering improved the prediction accuracy in most datasets

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

IV. PROBLEM IDENTIFICATION

After studying different literature we found that if we apply clustering data mining technique then we can efficiently categories lake level historic data. Hence forth we need an efficient clustering technique some author uses k-means for data clustering we found some drawbacks of k-means.

K-Means is a universally worn clustering algorithm for data mining. Clustering is a means of classifying n data objects into k clusters where every cluster has maximal resemblance as defined by an objective function. Every data object may only belong to one cluster, and the union of all clusters contains all n data objects.

- A key constraint of k-means is its cluster model. The notion is depend on spherical clusters those are distinguishable in a way in order that the mean value converges towards the cluster center. The clusters are probable to be of analogous size, in order that the assignment to the nearby cluster center is the right assignment. For example computing k-means clustering with a value of k=3 onto the any renowned for example Iris data set, the outcome repeatedly fails to split the three Iris species contained in the data set. With k=2, the two noticeable clusters (one contains 2 species) will be revealed, while with k=3 one of the two clusters will be separated into two uniform parts. In fact, k=2 is much apposite for this data set, in spite of the data set having 3 classes. As with several other clustering algorithms, the k-means result relies on the data set to gratify the assumption prepared by the clustering algorithms. It works well on a few data sets, while weakening on others.
- K-means having trouble with outliers.
- Often produce empty clusters as outcome.
- Though, K-means suffered from foremost shortcomings that have been a reason for it not being implemented on huge datasets. The most significant among these are K-means is sluggish and scales feebly with regard to the time it takes for huge number of data points.
- Complex to predict K-Value.
- It didn't work well with global cluster.
- Diverse initial partitions may outcome as different final clusters.
- It won't work fine with clusters (in the original data) of unlike size and dissimilar density.

V. METHODOLOGY

As we have discussed in problem identification section to overcome the drawback of k-means clustering algorithm we will use k-medoid algorithm accordingly our proposed scheme layout as follows:

Algorithm: K-Medoid

1. Initialize: randomly select(without replacement) *k* of the *n* data points as the medoids
2. Associate each data point to the closest medoid.
3. While the cost of the configuration decreases:
    1. For each medoid *m*, for each non-medoid data point *o*:
        1. Swap *m* and *o*, recompute the cost (sum of distances of points to their medoid)
        2. If the total cost of the configuration increased in the previous step, undo the swap

Multiple Regression: Multiple regression is an expansion of linear regression. It is used when we want to predict the value of multidimensional data points. The variable we want to forecast is known as dependent variable (or sometimes, the result, target or measure variable). The variables we are using to forecast the value of the dependent variable are known as independent variables (or sometimes, explanatory, the predictor). Multiple regression allows us to settle on the largely fit (variance explained) of the model and the comparative involvement of each one of the predictors to the total variance explained. For example, we might want to know how much of the difference in exam

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

performance can be explained by adjustment time, test apprehension, lecture attendance and gender "as a whole", but also the "relative contribution" of each independent variable in explaining the variance.

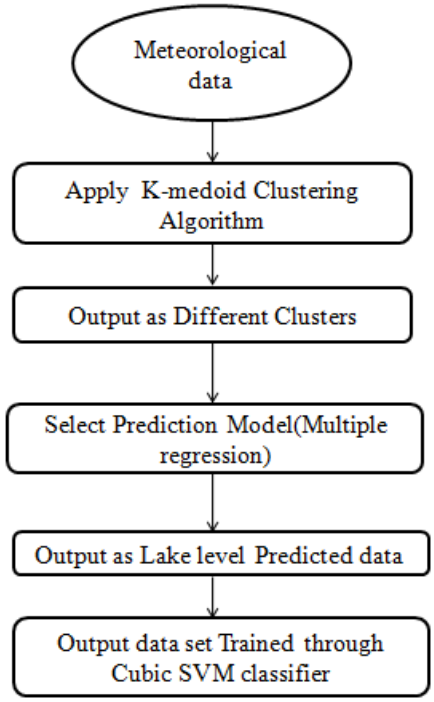

Fig.-2 Proposed Layout

General regression model as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- $\beta_0, \beta_1, \ldots, \beta_k$ are parameters
- $X_1, X_2, \ldots, X_k$ are known constants
- ε , the error terms are independent N(o, σ^2)

As we have drawn in proposed layout after applying multiple regression we will get predicted lake level value then further we will train this output data to Cubic SVM(Support Vector Machine) classifier so as to check the efficiency of forecasted data set.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

VI. RESULT AND DISCUSSION

For implementation of our project we have used Matlab 2015a and Meteorological Data taken form following Url.

URL:- http://lakepowell.water-data.com/

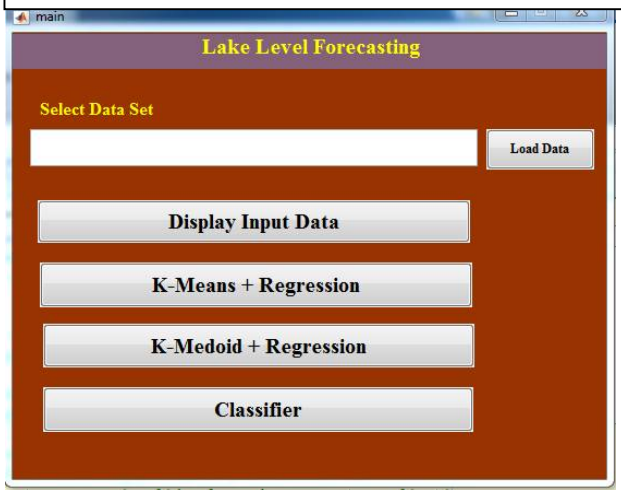

Fig.-3 Main GUI

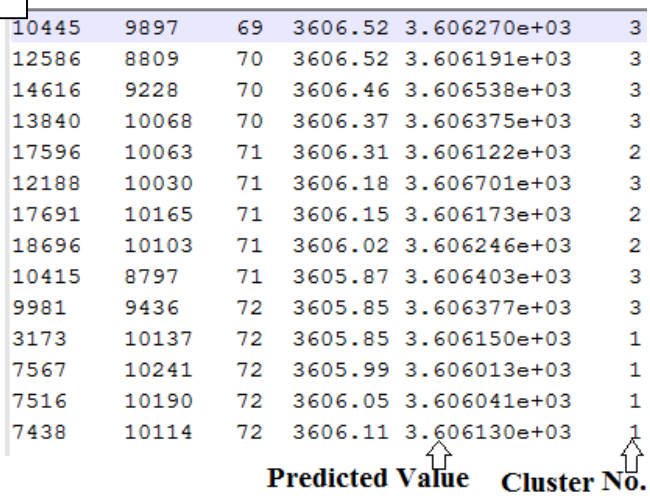

Fig.-4 Output of K-medoid Algorithm

Fig.-3 is the main GUI interface through which we will take input from user then clustering and regression performed over dataset. Fig.-4 is the snippets of output in which first column is Inflow, 2nd Column is Outflow, 3rd column is Water temperature, 4th column is Elevation of lake, 5th column is Predicted value of lake elevation and last column is cluster number.

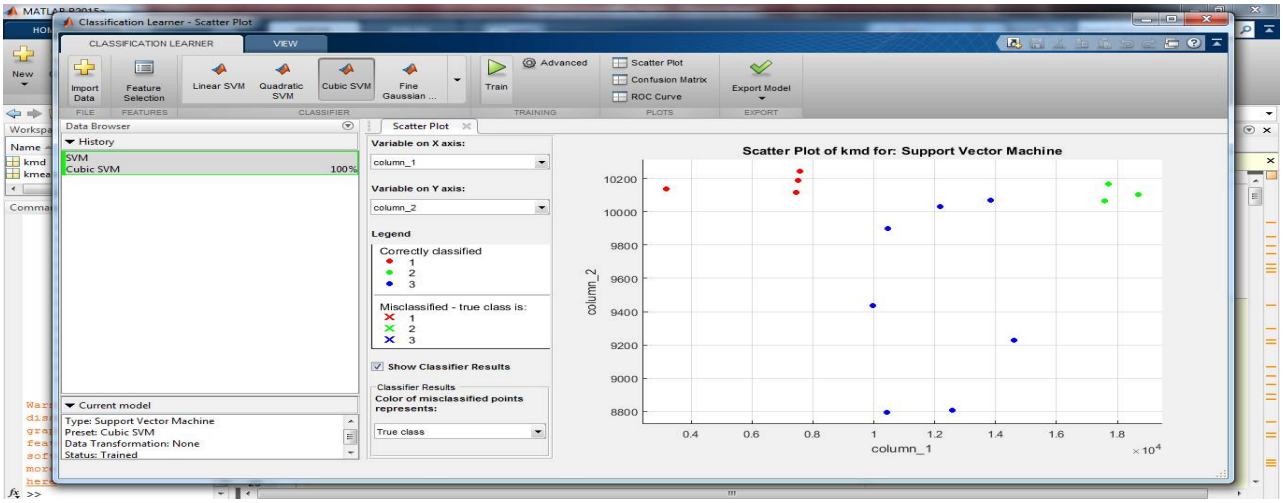

Fig.-4 Cubic SVM output after training (K-Medoid clustered data)

Fig.-4 shows the classifierlearner output in which predicted output data is passed to check the efficiency of predicted values. Learner creates confusion matrix which give value of TPR (True positive rate) and FNR (False Negative Rate) based these numeric values we have drawn graph which is displayed below in fig-5.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

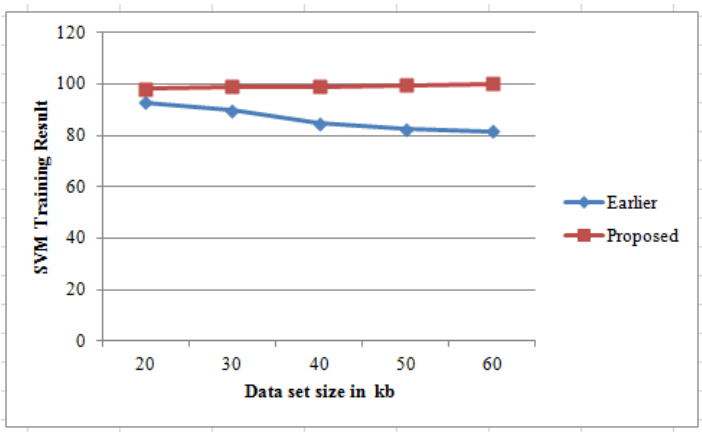

Fig.-5 Performance Comparison

VII. CONCLUSION

In this paper a portable and scalable prediction model is discussed for forecasting the water levels of lakes. Accurate predictions of water level fluctuation that results from hydro meteorological variations and anthropogenic disturbances are needed for sustainable development and management of lake water usage. Henceforth greater accuracy required for prediction system, we have compared k-means clustering algorithm and K-medoid clustering over Meteorological data and applied multiple regression prediction models over clustered data. Then to check the prediction efficiency we have passed predicted data to SVM classifier and found that K-Medoid produce better predicted value over k-means clustered data. In future we can increase the prediction accuracy by taking different regression model.

REFERENCES

[1] Yong Li, Yilin Wang, Hu Sheng, Feifei Dong, Rui Zou b, Lei Zhao c, Huaicheng Guo, Xiang Zhu, Bin He, "Quantitative evaluation of lake eutrophication responses under alternative water diversion scenarios: A water quality modeling based statistical analysis approach", Elsevier, 2013.
[2] Ozgur Kisi , Jalal Shiri , Sepideh Karimi , Shahaboddin Shamshirband , Shervin Motamedi , Dalibor Petkovic ´f, Roslan Hashim, "A survey of water level fluctuation predicting in Urmia Lake using support vector machine with firefly algorithm", Elsevier, 2015.
[3] Chih-Chieh Young, "Predicting the Water Level Fluctuation in an Alpine Lake Using Physically Based, Artificial Neural Network, and Time Series Forecasting Models", Hindawi Publishing Corporation, 2015.
[4] P. Delaney, "Integration of real-time monitoring, modelling and Forecasting for the development and operation of the Great lakes storm surge operational system Montreal", Quebec, 2015.
[5] C. J. Watras, "Decadal oscillation of lakes and aquifers in the upper Great Lakes region of North America: Hydroclimatic implications", AGU Publication, 2014.
[6] Ms. Ashwini Mandale, Mrs. Jadhawar B.A., "Weather forecast prediction: a Data Mining application", IJERGS, 2015.
[7] M.Viswambari, Dr. R. Anbu Selvi, "Techniques to Predict Weather: A Survey", International Journal of Innovative Science Engineering & Technology, Vol. 1, Issue 4, 2014.
[8] Folorunsho Olaiya, "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", MECS, 2012.
[9] T V Rajini Kanth, V V Sss Balaram and N. Rajasekhar Dhinaharan Nagamalai, "Analysis Of Indian Weather Data Sets Using Data Mining Techniques", Cs & It-Cscp, pp. 89-94, 2014.
[10] S. Gokila1, K. Ananda Kumar and A. Bharathi, "Modified Projected Space Clustering Model on Weather Data to Predict Climate of Next Season", Indian Journal of Science and Technology, Vol 8(14), 2015.
[11] Andrew d. Gronewold and Craig A. Stow noaa, "Unprecedented seasonal Water level dynamics on one of the earth'S largest lakes", Great Lakes environmental Research, 2014.
[12] Anne H. Clites, Joeseph P. Smith, "Water Levels Surge on Great Lakes", Elsvier, 2015. Guerard, J.B., Jr. and E. Schwartz, "Regression Analysis and Forecasting Models", 2007.
[13] P.Hemalatha, "Implementation of Data Mining Techniques for Weather Report Guidance for Ships Using Global Positioning System", International Journal of Computational Engineering Research, Vol. 3, Issue 3, 2013.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

[14] Prashant Shrivastava, S. Pandiaraj and Dr. J. Jagadeesan, "Big Data Analytics In Forecasting Lakes Levels", International Journal of Application or Innovation in Engineering & Management, Vol. 3, Issue 3, 2014.
[15] N. Sivaram. K. Kumar, "Applicability of Clustering and Classification Algorithms for Recruitment Data Mining", International Journal of Computer Applications, Vol. 4, No.5, 2010.
[16] Shubhendu Trivedi, Zachary A. Pardos and Neil T. Heffernan Department of Education IES Math Centre for Mathematics and Cognition grant. "The Utility of Clustering in Prediction Tasks", 2011.
[17] Moslem Imani "Forecasting Caspian Sea level changes using satellite altimetry data (June 1992–December 2013) based on evolutionary support vector regression algorithms and gene expression programming", Elsevier, 2014.
[18] Hakan Tongal, R. Berndtsson, "Phase-space reconstruction and self-exciting threshold modeling approach to forecast lake water levels", Elesvier, 2015.
[19] F. Hossain, A. H. Siddique-E-Akbor, L. C. Mazumder, S. M. Shahnewaz, S. Biancamaria, H. Lee, C. K. Shum, "Proof of Concept of an Altimeter-Based River Forecasting System for Trans boundary Flow Inside Bangladesh", 2014.
[20] Kalyan city blogspot.com/what is forecasting, meaning, features and limitations.
[21] K. H. V. Durga Rao et. al. Kedarnath flash floods: a hydrological and hydraulic simulation study, Elsevier, 2014.