

Real Time Prediction System of Discharge of the Rivers using Clustering Technique of Data Mining

Anamika Gupta¹, Dr. Setu Kumar Chaturvedi²

¹Research scholar, Technocrats Institute of Technology, Bhopal

²Professor and Head of Comp. Sci. & Engg. Dept. Technocrats Institute of Technology, Bhopal.

Abstract:- In this paper, the investigation is carried-out in hydrological daily discharge time series of the Panchratna station in the Brahmaputra River, in India. The values of discharge of 5 years have been taken. Statistical parameters (mean, max, average, range and standard deviation) of the data are calculated and standardized using z-score method. For this work, agglomerative Hierarchical Clustering (AHC) and K-Means clustering of Time Series Data Mining (TSDM) technique have been used. By using these clustering techniques discharge patterns for the next year have been predicted. At first, patterns for every year (all 5 years, taken) have been generated and after that by using both clustering techniques discharge patterns for the next year have been predicted and a comparative analysis has been done between both the techniques. The coefficient of determination (R^2) is used as the criteria to evaluate the adequacy of the result. For this work, only monsoon duration (june-sep) has been taken for the prediction of discharge pattern. By predicting discharge patterns, flood forecasting can be done more easily and accurately and a big loss of life and property can be saved.

Keywords:- Data-mining, Discharge, Prediction System, Hydrological Process, Clustering Techniques.

I. INTRODUCTION

Data mining refers to extracting or mining knowledge from large amounts of data. The Time Series Data Mining (TSDM) methodology follows the time delayed embedding process to predict future occurrences of important events. TSDM framework combines the methods of phase space reconstruction and data mining to reveal hidden patterns predictive of future events in nonlinear, nonstationary time series. Time series data mining is dedicated to the development and application of novel computational techniques and patterns for the analysis of large temporal databases. Time series is an important class of temporal data objects as well as it can be easily obtained from scientific and financial applications. The time series data is large in data size, high dimensional and necessary to update continuously. The rapid development of data mining provides a new method for water resources management, hydrology and hydro informatics research. In hydrology, data mining depends on the hydro-meteorological data, which generally takes the structure of time series. Hydrological time series are the sets of various record values of hydrological data that diverge with time. Researches which are based on data mining theory and hydrological techniques are needed to analyse hydrological daily gauge, rainfall, sediment, evaporation, temperature and discharge time series of the particular river in particular station for various types of study.

II. LITERATURE REVIEW

A review of relevant literatures has been done as part of this work. The focus of the review has been to identify sources of information and techniques that would help to develop a methodology for the prediction of discharge as well as flood forecasting. The various researches have been done in hydrological forecasting using different techniques of data mining and artificial Intelligence like time Series data mining, artificial neural network, rough set theory and support vector machine.

Data mining is categorized into directed and undirected data mining. Clustering, association rules, description and visualization are the techniques used in undirected data mining [10]. Clusters are defined as a collection of data objects that are similar to each other within the same cluster and are dissimilar to the objects in other clusters [1]. Time series forecasting approach of Neural Network has been applied with Multi resolution Learning Paradigm (NNMLP) to forecast stream flows over a long term range (over 12 months), to improve long-term range stream flow forecast [11]. A novel hybrid forecasting model known as GLSSVM has been proposed, which combines the Group Method of Data Handling (GMDH) and the Least Square Support Vector Machine(LSSVM). GMDH and LSSVM Models used the long term observations of monthly river flow discharge [7]. A long-term integrative ANN runoff forecasting model is built up, data mining technology is applied to hydrologic forecasting and Back Propagation algorithm of ANN is used to make the data mining work. This approach has less input data requirement, less maintenance and performs more simple forecasting

process, as compare to traditional runoff forecasting method [5]. A novel similarity analysis method based on BORDA count for Multivariate Time Series (MTS) similarity mining has been given. It visualizes the results to users more directly [8]. Subsequence time series (STS) Clustering Techniques for meaningful pattern discovery has been discovered [4]. The single element, medium and long term classification forecast model using historical data of a hydrologic series of forecast factors has been presented which was based on rough set theory [9]. On the issues of forecast of surface runoff a method has been given which was based on ANN technology [3]. Using wavelet neural network model, monthly runoff forecasting has been done [2].

On the basis of above literature survey, the time series data mining techniques is opting for our purpose. K-Means Clustering, Dynamic time warping algorithms, Agglomerative Hierarchical Clustering and wards criterion proposed for pattern discovery.

III. PROPOSED METHODOLOGY

3.1 CLUSTERING TECHNIQUE

Clustering is a technique of data mining for grouping objects based on the information found in the data describing the objects or their relationships. This means that the objects in a group will be similar to one other and different from the objects in other groups. Clustering discovers remarkable patterns in the underlying data. The greater similarity (or homogeneity) within a group, and the greater the difference between groups, the “better” or more distinct the clustering will be [1]. In this research, K-Means and AHC (Agglomerative Hierarchical Clustering) clustering techniques have been used.

3.2 PROPOSED ALGORITHM

Hydrological Process Algorithms

Hydrol Process

1. Select a data set
2. Find statistical parameters
3. Standardize the data
4. Apply K means clustering and find out number of clusters
5. Apply DTW and find out similarities and dissimilarities
6. Apply AHC algorithms and find out dendograms
7. Calculate moving average in monthly discharge standardized data
8. Find patterns
9. Validation of results based on causal effect

3.2.1 Data Set Selection and Formation of Data

For this work, daily discharge, water level and rainfall data from CWC (central water commission) and meteorological websites have been collected. The water level data is in hourly format and discharge data on daily basis. As per our requirement, the discharge data has been converted into monthly discharge, for better results. The rainfall data is measured twice a day viz. morning and evening. Firstly, the rainfall data is converted into daily basis and then in monthly average form.

3.2.2 Statistical Analysis and Standardization of Data

The four statistical characteristics (Q_{mean} , Q_{max} , Q_{range} , $Q_{dev.}$) were obtained for each month in discharge data as formula is given in table3.1. There were five years accounted viz. 2003,2004,2005,2006,2007. In order to have an effective analysis the data were standardized using Z-scores technique so that the mean of the entire data range leads to 0 and the standard deviation is 1. The need for standardization was that to avoid affecting the study results by the wide variations in the data. The reason for such preference is that calculating z requires the discharge mean and the discharge standard deviation, not the sample mean or sample deviation.

| | |
|------------------------------|--|
| Mean monthly discharge | $Q_{\text{mean}} = (1/n)\sum_{i=1}^n x_i$ |
| Monthly maximum discharge | $Q_{\text{max}} = \text{Max}(x_1, x_2, \dots, x_n)$ |
| Monthly minimum discharge | $Q_{\text{min}} = \text{Min}(x_1, x_2, \dots, x_n)$ |
| Monthly discharge range | $Q_{\text{range}} = Q_{\text{max}} - Q_{\text{mean}}$ |
| Standard deviation discharge | $Q_{\text{std}} = (1/n)\sum_{i=1}^n (x_i - Q_{\text{mean}})^2$ |

Table 3.1: Statistical Calculations

3.2.3 Apply K-Means Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. Among all clustering algorithms, K-Means clustering is more efficient for finding spherical-based clusters capability in small- to medium-sized databases.

Algorithm: The K-Means algorithm for partitioning, where every cluster's center is represented by the mean value of the objects within the cluster.

Input:

'k' is the number of clusters,

'D' is a data set containing 'n' objects.

Output: A set of k clusters.

Method:

1. Arbitrarily select k objects from D as the initial cluster centers;
2. repeat
3. (re)assign every object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. update the cluster means, i.e., calculate the mean value of the objects for every cluster;
5. until no change;

3.2.4 Segmentation of Data

In this, the hydrological period is divided into 3 segments as pre-monsoon(Jan-May), monsoon(Jun-sep) and post-monsoon(Oct-Dec). The statistical attributes obtained were subjected to K-Means clustering. According to the cases distribution in the clusters, the annual discharge process could be obtained as separate 3 clusters segments. It can be appropriately inferred that the highest discharge occurs in the monsoon periods.

3.2.5 Similarity Measures

Similarity search in time series analysis is one of the most rapidly developing and demanding fields in data mining. Unlike normal database queries, which find data that match the given query exactly, a similarity search finds data sequences that differ only slightly from the given query sequence. It can be classified into two categories:

Whole matching: In this kind of matching the time series data has to be of equal length.

Subsequence matching: In this mentioned category of matching, a query sequence X and a longer sequence Y have been taken. The objective is to identify the subsequence in Y, beginning at Y_i , which best matches X, and report its offset within Y. The primary difficulty is defining a similarity measure [6]. For similarity analysis of time series data, Euclidian distance is typically used as a similarity measure.

Given two sequences

$X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_m)$ with $n = m$, their Euclidean distance is defined as follows:

$$D(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

DTW (Dynamic Time Warping) Algorithm: is an algorithm for measuring optimal similarity between two time data sequences. The time series data varies not only on the time amplitudes but also in terms of time progression as the hydrological processes may reveal with different rates in response to the different environmental conditions. A non-linear alignment produces a similar measure, allowing similar shapes to match even if they are out of phase in time axis [12]. The sequences are "warped" non-linearly in the time dimension

to determine a measure of their similarity independent of certain non-linear variations in the time dimension. To find the best alignment between time sequences X&Y one needs to find the path through the grid.

3.2.6 Apply AHC Algorithm

To identify the discharge pattern from the corresponding discharge time series data, each hydrological period obtained after the segmentation from the k- means clustering has been taken and then analysis of the discharge pattern in each of the periods has been done. The analysis involved, hierarchical clustering techniques on the 5 years as the cases and the attributes, as observations of the average discharge data for the months which are in the hydrological period. Now the discharge time series data of the cluster center is the pattern because all other objects in a particular group then show similarity to the center only. Thus the cluster center can be taken as the pattern of discharge.

Algorithm: Agglomerative Hierarchical Clustering, an unsupervised learning technique, is a subdivision of hierarchical clustering and a bottom up approach, which proceeds by series of fusions of the N objects into groups. If it is given a set of N items to be clustered and an N*N distance (or similarity) matrix then the basic process of agglomerative hierarchical clustering can be done iteratively following these four steps:

1. Start with N clusters each containing a single entity, and an $N \times N$ symmetric matrix of distances (or similarities) Let d_{ij} = distance between item i and item j.
2. Search the distance matrix for the nearest pair clusters (i.e., the two clusters that are separated by the smallest distance). Denote the distance between these most similar clusters U and V by d_{UV} .
3. Merge clusters U and V into new clusters labeled T. Update the entries in the distance matrix by
(a) Deleting the rows and columns corresponding to clusters U and V, and
(b). Adding a row and column giving the distances between the new cluster T and all the remaining clusters.
4. Repeat steps (2) and (3) a total of N-1 times.

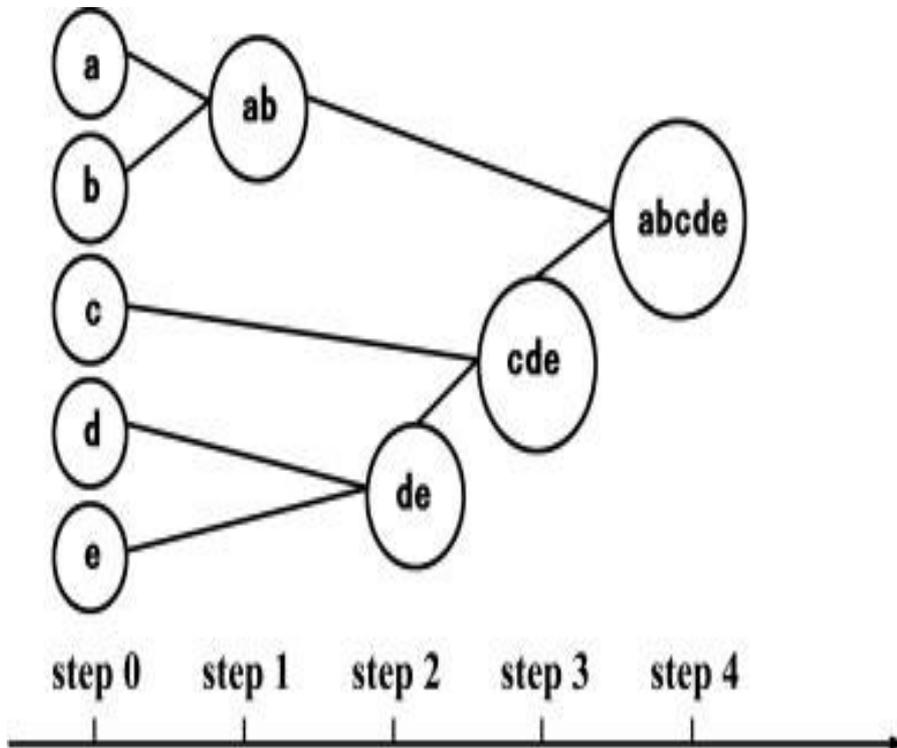


Fig 3.1: Cluster tree obtained by AHC

In our analysis, the discharge pattern (time series data of discharge for the months in the hydrological period) of a year (among 5 years) is clustered into several clusters. But the year which formed the cluster center formed the pattern with its discharge data for the months in the period. All other members (years) in the cluster attained membership of the cluster because there was similarity to the year representing the center so they can be said to follow the pattern.

- Now, the center (the year) in the cluster is obtained, plotting of the discharge data of that year corresponding to daily discharge in the months, along the x-axis would give the pattern.

Specifications taken during HC application:

- Proximity type: Dissimilarities
- Distance: Euclidean
- Agglomeration method: Ward's
- Cluster: along rows

3.2.7 Calculate Moving Average in Monthly Discharge Standardized Data

A moving average, also called rolling average is a type of finite impulse response filter used to analyse a set of data points by creating a series of averages of different subsets of the full data set. A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. The threshold between short-term and long-term depends on the application, and the parameters of the moving average will be set accordingly.

The formula of simple moving average is given below:

$$s_t = \frac{1}{k} \sum_{n=0}^{k-1} x_{t-n} = \frac{x_t + x_{t-1} + x_{t-2} + \dots + x_{t-k+1}}{k} = s_{t-1} + \frac{x_t - x_{t-k}}{k}, \quad (2)$$

S_t =simple moving average

K =observations

t =data point with respect to time

n =no. of data points

3.2.8 Find Patterns

At last, the standardized discharge data of monsoon periods (from June - September) is taken and moving average of monthly discharge standardized data is calculated, to find out the patterns by using it.

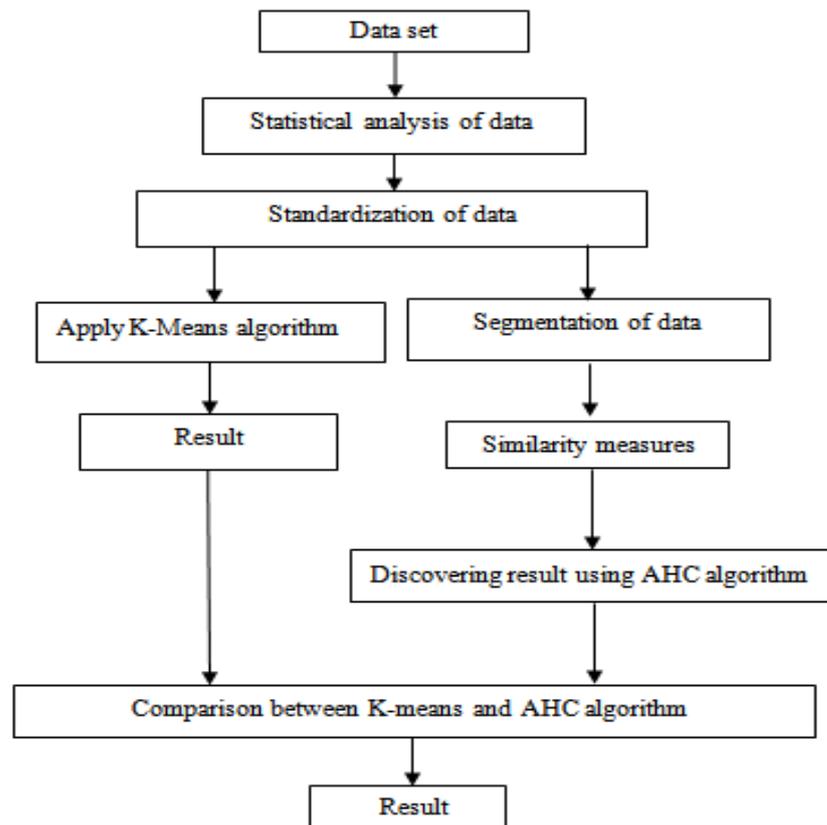


Figure 3.2: Proposed System

IV. EXPERIMENTAL SETUP

4.1 LOCATION SELECTED

For this work river Brahmaputra was selected due to highly flood prone area. The site Panchratna (Latitude 26° 11' 55" and Longitude 90° 34' 38") of the river Brahmaputra, located in the district of Goalpara in the state of Assam is selected for the study. The length of the river upto the site is 2562 Km. The catchment area upto the site measures 468790 sq km. The site is located on the left bank of river. The type of site is HO (Hydrological Observation). For this work, daily discharge and water level data for the entire year were taken during the highly flooded years of 2003, 2004, 2005, 2006 and 2007. The average max temperature is 30°C and minimum temperature is 15°C, recorded during the Pre Monsoon period in the Goalpara. Average humidity percentage is recorded as 82% during the months from October to December.

4.2 EXPERIMENT DESIGN

To conduct experiments XLSTAT, a data mining tool, is used which contains implementation of various clustering algorithms like K-mean, AHC etc. and other data mining techniques. For this work daily discharge data of monsoon duration viz. june to September of 5 years, is used. And then by using this tool we standardize our data at first. In this study XLSTAT software is used for finding clusters by using K -Means clustering and dendograms by using AHC algorithm etc. XLSTAT software is compatible with all Excel versions from version 97 to version 2013 (except 2008 for Mac). Also it is compatible with the Windows 9x till Windows 7 systems, and with the Power PC as well as Intel based Mac systems.

4.3 EVALUATION

The coefficient of determination (R^2) is used as the criteria to evaluate the adequacy of result. The coefficient of determination based on the flow forecasting errors is calculated as,

$$R^2 = 1 - \frac{\sum_{i=1}^n (F_{i(flow)} - F_{i(model)})^2}{\sum_{i=1}^n (F_{i(flow)} - F_{mean})^2} \quad (3)$$

where n is the number of observed data, $F_{i(flow)}$, $F_{i(model)}$ and F_{mean} are monthly flow measurement, the results of developed flow model and mean flow measurements, respectively[13].

V. RESULT AND ANALYSIS

5.1 RESULTS OF K-MEANS ALGORITHM

The hydrological period is divided into 3 segments as pre- monsoon(Jan-May), monsoon(June-Sep) and post-monsoon(Oct-Dec). The statistical attributes obtained were subjected to K-means clustering. A total of 60 (12*5=60) cases based on 4 attributes from data of 5 years can be used. The specifications taken during clustering were clustering criterion (Determinant), Initial Clusters (Random Generated), Repetitions 10. According to the cases distribution in the clusters, the annual discharge process could be obtained as separate 3 clusters as premonsoon, monsoon and post monsoon. For this work, only the Monsoon period has been taken. So we have taken 20(4*5=20) cases based on 4 attribute from the data of 5 years and it can be aptly inferred that the graph shows the highest discharge in the monsoon periods.

5.1.1 Clusters Observation

By applying, K-means clustering algorithm on the data set, five different clusters have been found.

| Observation | Cluster | Distance to centroid |
|-------------|---------|----------------------|
| 1-Jun-03 | 1 | 1127.308 |
| 1-Jul-03 | 2 | 5949.229 |
| 1-Aug-03 | 3 | 9311.814 |
| 1-Sep-03 | 4 | 7541.416 |
| 1-Jun-04 | 1 | 12029.560 |
| 1-Jul-04 | 2 | 12000.475 |
| 1-Aug-04 | 1 | 5326.881 |
| 1-Sep-04 | 1 | 12577.838 |
| 1-Jun-05 | 2 | 6547.026 |
| 1-Jul-05 | 5 | 2371.099 |

| | | |
|-----------------|---|----------|
| 1-Aug-05 | 5 | 4733.451 |
| 1-Sep-05 | 4 | 6393.913 |
| 1-Jun-06 | 4 | 9385.251 |
| 1-Jul-06 | 3 | 9311.814 |
| 1-Aug-06 | 5 | 3807.807 |
| 1-Sep-06 | 2 | 4387.643 |
| 1-Jun-07 | 1 | 3071.894 |
| 1-Jul-07 | 2 | 4409.076 |
| 1-Aug-07 | 2 | 9285.587 |
| 1-Sep-07 | 2 | 1260.790 |

Table 5.1: Cluster Observations after Applying K-Means

5.1.2 Cluster analysis of K-Means:

Cluster centroids of the clusters have been calculated and after that a profile plot, which is a graphical representation of the clusters, of the clusters has been plotted by applying K-Means algorithm.

| Cluster | Y1 | Y2 | Y3 | Y4 |
|----------------|-----------|-----------|-----------|-----------|
| 1 | 34607.800 | 26589.538 | 4403.255 | 14387.600 |
| 2 | 45017.286 | 30990.636 | 8088.922 | 25421.000 |
| 3 | 77540.500 | 49828.175 | 13442.990 | 38411.000 |
| 4 | 62238.333 | 35256.390 | 14430.973 | 45792.667 |
| 5 | 51069.667 | 45250.473 | 4046.351 | 12966.333 |

Table 5.2: Cluster Centroid of the Clusters after Applying K-Means

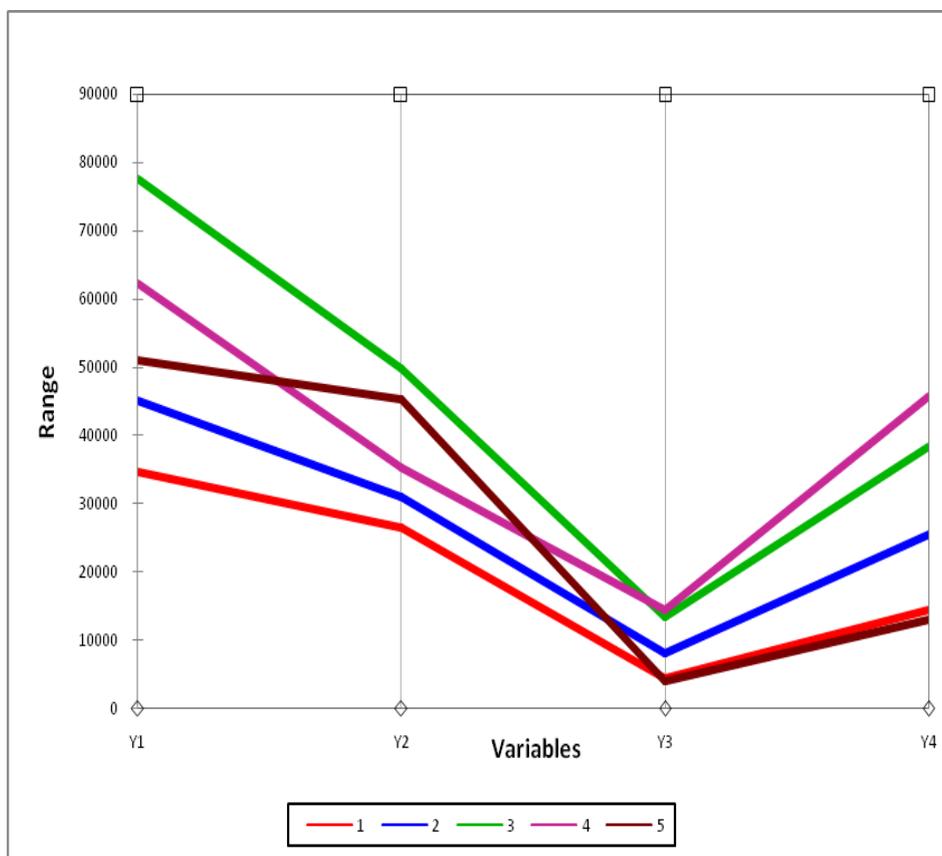


Fig5.1: Graphical Representation of Clusters obtained by applying K-Means

5.2 RESULTS AFTER APPLYING AHC ALGORITHM

5.2.1 Clusters Observations

By applying, Agglomerative Hierarchical Clustering (AHC) algorithm on the data set, three different clusters have been found.

| Observation | Cluster |
|-------------|---------|
| 1-Jun-03 | 1 |
| 1-Jul-03 | 1 |
| 1-Aug-03 | 2 |
| 1-Sep-03 | 2 |
| 1-Jun-04 | 1 |
| 1-Jul-04 | 1 |
| 1-Aug-04 | 1 |
| 1-Sep-04 | 1 |
| 1-Jun-05 | 1 |
| 1-Jul-05 | 3 |
| 1-Aug-05 | 3 |
| 1-Sep-05 | 2 |
| 1-Jun-06 | 2 |
| 1-Jul-06 | 3 |
| 1-Aug-06 | 3 |
| 1-Sep-06 | 1 |
| 1-Jun-07 | 1 |
| 1-Jul-07 | 1 |
| 1-Aug-07 | 1 |
| 1-Sep-07 | 1 |

Table 5.3: Cluster Observations after Applying AHC

5.2.2 Cluster analysis of AHC:

Cluster centroids of the clusters have been calculated and after that a profile plot of the clusters has been plotted by using these values of cluster centroids by AHC algorithm.

| Cluster | Var1 | Var2 | Var3 | Var4 |
|---------|-----------|-----------|-----------|-----------|
| 1 | 40680.000 | 29156.845 | 6553.227 | 20823.750 |
| 2 | 66918.750 | 36969.340 | 14809.278 | 44704.000 |
| 3 | 56832.500 | 48324.895 | 5770.211 | 18570.750 |

Table 5.4: Cluster Centroid of the Clusters after Applying AHC

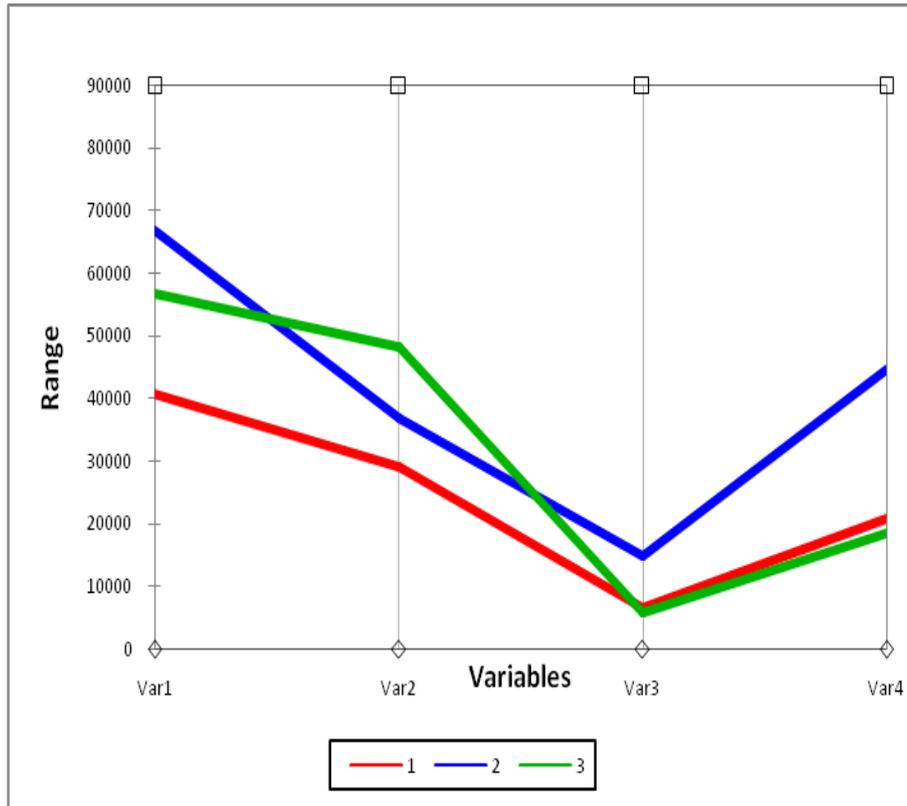


Fig5.2: Graphical Representation of Clusters obtained by applying AHC

5.2.3 Dendograms obtained by applying AHC

AHC is a bottom-up approach of hierarchical clustering, which proceeds by series of fusions of the N objects into groups. In the fig 5.3 the clusters are merging and converting into the final cluster.

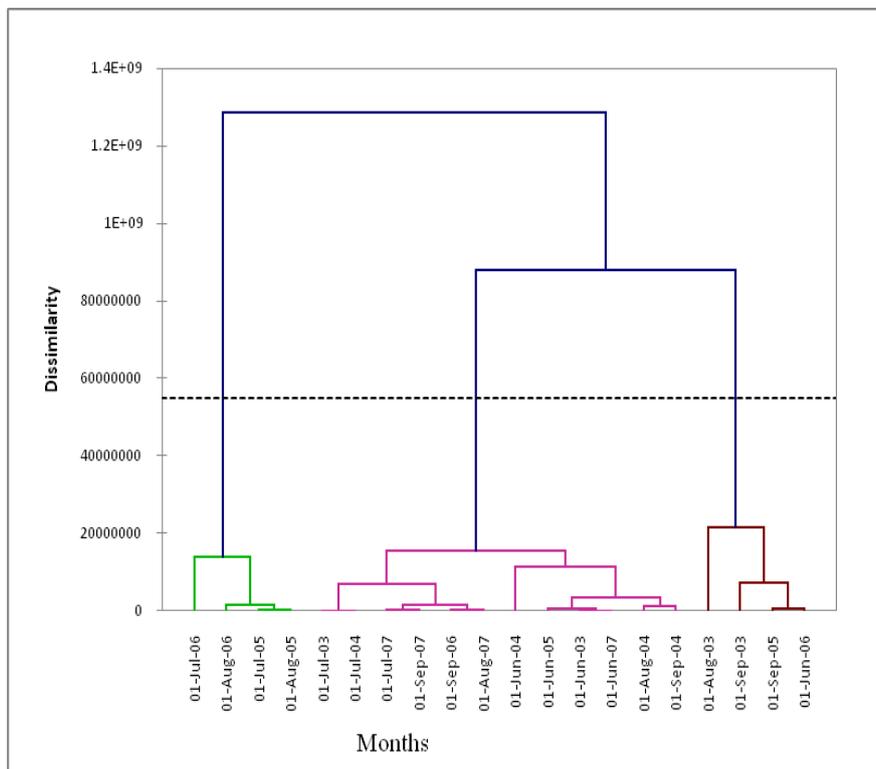


Fig 5.3: Clusters Merging and Conversion into the Big Cluster

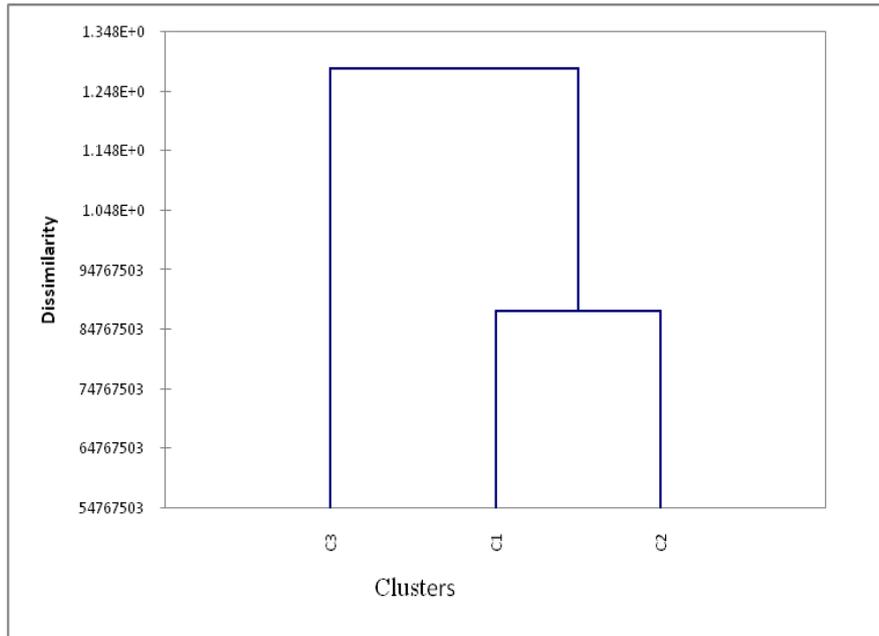


Fig 5.4: Final Clusters obtained after Merging of Sub-clusters

5.3 Validation of Results based on Causal Effect

It is tried to analyse the nature of variation of the water level of the river during the monsoon period, in this work. For this a hydrograph plotting the water level data along y-axis and the days taken at an interval of 7days from the month of June to September (Monsoon period) is obtained. The hydrological processes showed a causal-effect relationship with several happenings in nature as well as human interferences. As it is inferred that the annual discharge processes is mainly affected by the rainfall and the temperature too. The rainfall distribution was taken for 5 years. Our focus of study was the monsoon period of June-September. The graphs were plotted showed similar pattern that conformed to the discharge pattern, which shows that rainfall was related strongly as a contributor to discharge. The distribution of rainfall could be seen throughout for the period of monsoon. In this work, the following observations strongly validate our obtained patterns:

- 1) The discharge patterns are similar to the corresponding year hydrographs of water level during the monsoon period.
- 2) The discharge patterns are in rising trend during the highly rainfall periods of the monsoon.
- 3) The rainfall - runoff graph have more similarities.

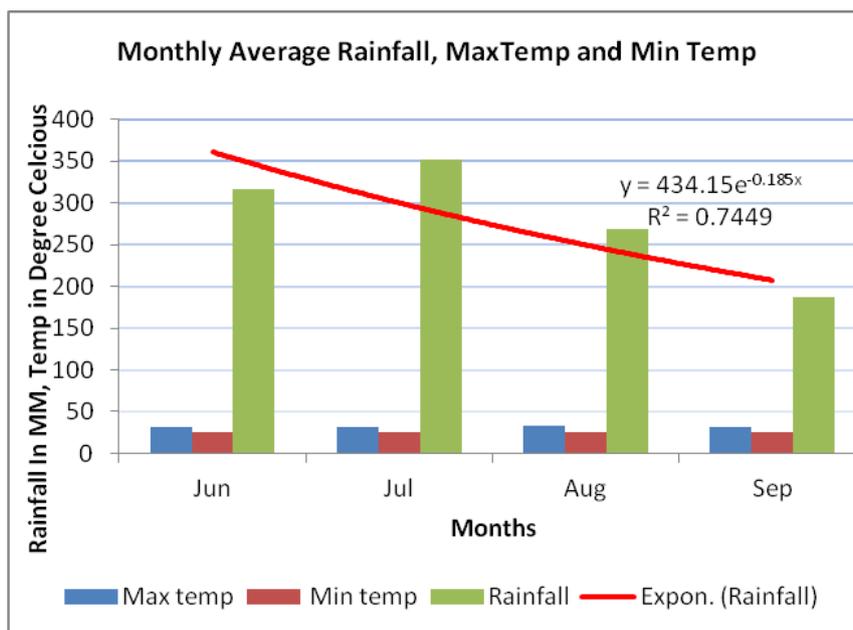


Fig 5.5: Validation of Results Based on Causal Effect

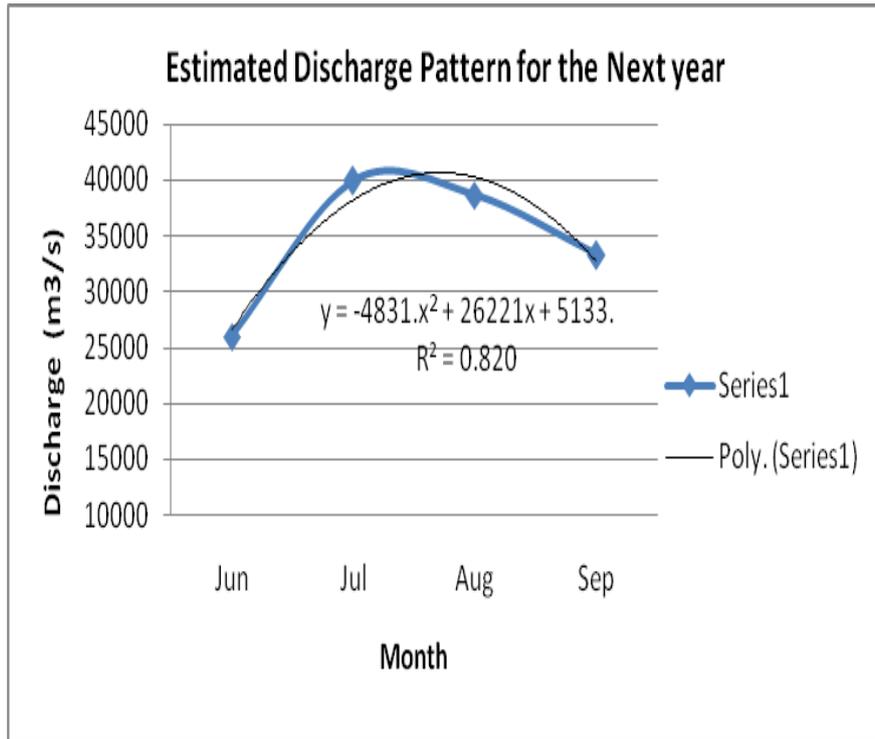


Fig 5.6: Predicted Pattern of Discharge (using AHC)

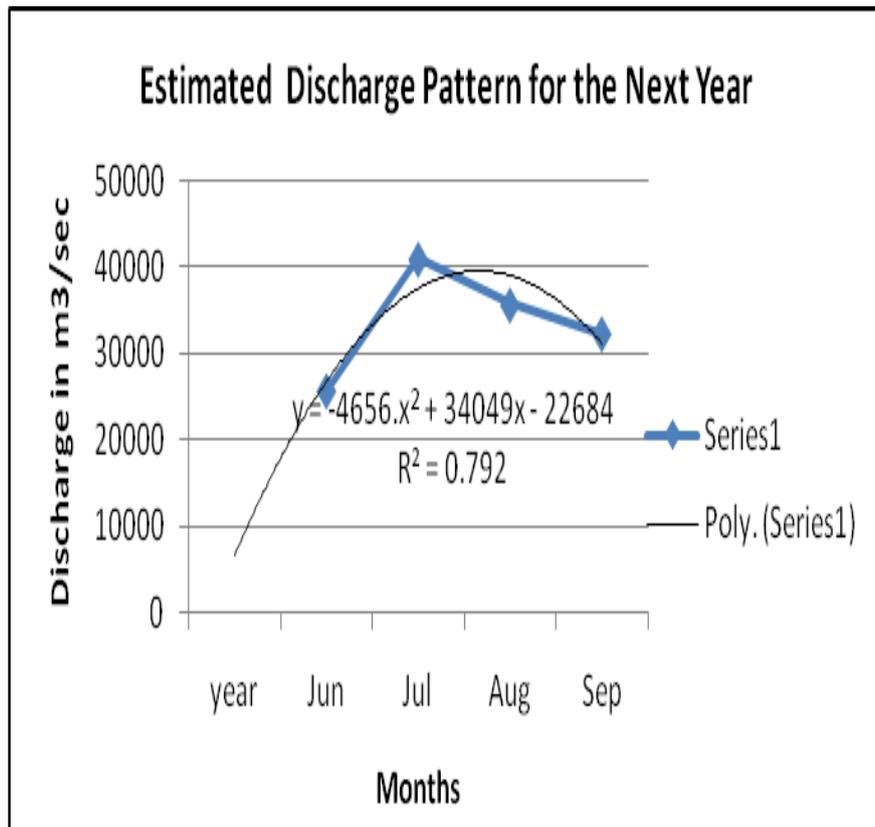


Fig 5.7: Predicted Pattern of Discharge (using K-means)

5.4 COMPARATIVE RESULT ANALYSIS

In this research two clustering techniques i.e. K-Mean and agglomerative hierarchical clustering (AHC) are applied, to find out the result. The comparison between the results is as follows.

| K-Means | AHC |
|---|---|
| The number of clusters obtained after applying k-means algorithm is 5. | The number of clusters obtained after applying AHC algorithm is 3. |
| After applying k-means, the value of R^2 (coefficient of determination) obtained, is 0.792. | After applying AHC, the value of R^2 (coefficient of determination) obtained, is 0.820. |
| It's a partitioning clustering method. | It's a hierarchical clustering method. |
| This is a simple approach. | AHC method is a more robust approach. |

Table 5.5: Comparison between Clustering Techniques

After applying K-Means and agglomerative hierarchical clustering techniques, it is find out that agglomerative hierarchical clustering technique is more accurate.

VI. CONCLUSION AND FUTURE WORK

In this work, data mining technique like clustering, similarity search and pattern discovery were used in hydrological discharge time series data and discovered the results as a similar discharge process and patterns. The comparison of hydrographs and rainfall during the same time period has been done and it is proved that the discharge patterns were more similar under the same climatic periods. The patterns found by the AR Model can be used for the prediction of future value of discharge.

In Indian continent the whole river system is divided in three periods viz pre monsoon, monsoon and post-monsoon period. For this work, only monsoon data during the high floods year is taken. The future work will focus on post monsoon and pre monsoon data which will show the complete study of hydrological behaviour of river during high floods year for the particular stations.

REFERENCES

- [1]. Han, J.W., and Kamber, M., Data Mining Concept and techniques, Morgan Kaufman: San Fransciko, CA, Issue 7, pp. 78-90, 2001.
- [2]. Aiyun, L., and Jiahai, L., 'Forecasting monthly runoff using wavelet neural network model, International conference on Mechatronic Science', Electronic Engineering and Computer, IEEE, 978-1-61284-722-1, 2011.
- [3]. Feng, L.H., Zhang, J.Z., 'Application of ANN in Forecast of surface runoff, Networked Computing (INC)', 2010 6th International Conference on, E-ISBN: 978-89-88678-20-6, pp.1-3, 2010.
- [4]. Kadir, A. Peker, 'Subsequence Time Series (STS) Clustering Techniques for Meaningful Pattern Discovery', KIMAS Waltham, MA, USA, IEEE,0-7803-9013,pp.78-90, 2005.
- [5]. Li, C., and Yuan, X., 'Research and Application of Data Mining for Runoff Forecasting, Intelligent Computation Technology and Automation (ICICTA)', Volume:1, pp. 795-798, 2008.
- [6]. Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E., 'Searching and mining trillions of time series sub sequences under dynamic time warping', ACM, 978-1-4503-1462, pp.340-356, 2012.
- [7]. Samsudin, R., Saad, P., and Shabri, A., 'River flow time series using least square support vector machines', Hydrol Earth Syst Sci, 15, pp.1835-1852, 2011.
- [8]. Yuelong, Z., Shijin, L., Dingsheng, W. and Xiaohua, Z., 'A Novel Approach to the Similarity Analysis of Multivariate Time series and its Application in Hydrological Data mining', International Conference on Computer Science and Software Engineering, IEEE, Vol. 4, pp. 730-734, 2008.
- [9]. Sihui, D. 'Forecast Model of Hydrologic Single, Element Medium and Long-Period Based on Rough Set Theory, Fuzzy Systems and Knowledge Discovery', Sixth International Conference, Volume: 1, pp. 19-25, 2009.
- [10]. Berry, M. J. A., Linoff, G., 'Mastering Data Mining', John Wiley & Sons, Wiley: 1 edition (December 28, 1999), ISBN-10, pp. 1-512, 2000.
- [11]. Liang, X., and Liang, Y., 'Application of data mining in Hydrology', ICDM 2001, pp. 617-620, 2001.

- [12]. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. J., 'Querying and mining of time series data: experimental comparison of representations and distances measures', PVLDB,1,2, 1542-52, 2008.
- [13]. Terzi, O., 'Monthly River Flow Forecasting by Data Mining Process', Suleyman Demirel University, Turkey, Sciences Journal des Sciences Hydrologiques, 51, ISBN 978-953-307-154-1, pp. 588-598, 2011.