Association Rules of Data Mining for the Characteristic Analysis of Sub-basins of a River

Dr.Kirti Malhotra
Professor and Head, Department of civil Engg. Acharya Institute of Technology, Bangalore-90

H.Venugopal
Professor, Department of Computer Science and Engg. Sri Siddhartha Institute of Technology, Tumkur-05

ABSTRACT

The sub basins of a river are not hydrologically homogeneous, because of their location, drainage pattern, precipitation and other characteristics. The present study is a new approach for developing relationships between different hydrological parameters such as cloud cover, potential evapotranspiration (PET), Reference Crop Evapotranspiration (RCET), vapor pressure, temperature, precipitation and discharge of different sub basins. The study considers the application of association rules of data mining for 8 sub basins of a river in south India. An attempt is also made to check whether the developed association rules in the data hyperspace have any physical meaning or not. The generated association rules indicate there is hydrological homogeneity between some sub basins while others are hydrologically heterogeneous.

General Terms

Data Mining, Association Rules, Classification and prediction,

Keywords

Cloud cover, Potential Evapotranspiration (PET), Reference Crop Evapotranspiration (RCET), Vapor Pressure, Temperature, Precipitation and Discharge.

1. INTRODUCTION

The demand for water is increasing constantly while the availability of water remains practically constant. Hence, proper development and management of available water resources is a necessity. Estimation of dependable yield is one of the primary inputs for the development and management of any water resource system. The dependable yield/runoff of a river basin/sub basin depends on rainfall characteristics such as magnitude, intensity, distribution and catchment characteristics such as soil, vegetation, shape, geology, slope and drainage density and also climatic factors which influence evapotranspiration. The interrelationships between these factors are extremely complex.

The analysis of sub basins' characteristics of a river is of vital interest in hydrological engineering due to its importance in estimating dependable yield for the design and management of water related projects. The analysis also is useful in flood and drought analysis, prediction and control.

2. DATA MINING

Data mining is considered as knowledge discovery in data bases.[Han,J. and Kamber,M. 2006] It is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases, data warehouses and other massive information repositories.

The abundance of hydrological data coupled with need for powerful data analysis tools has been termed as data rich but information-poor situation. Hence the important decisions such as prediction of rainfall, runoff including floods are made not based on information rich data stored in databases but rather on decision makers intuition simply because the decision maker may not have the tools to extract the valuable knowledge embedded in large amounts of data.

Available current expert system technologies rely on users or domain experts to manually input knowledge into the knowledge databases.[Piatetsky-Shapiro,G. 1991] But this procedure is prone to biases and errors and is extremely time consuming and costly. Artificial Neural Networks(ANNs) are useful for clustering the river basins on the basis of hydrological homogeneity.[Thandaveswara, et al., 2000]

Data mining tools perform data analysis and may uncover important data patterns contributing greatly to strategic decisions, knowledge bases, scientific and medical research including hydrology [Nagesh kumar,D. and Dhanya,C.T.2009]. Time series Data mining which combines chaos theory and Data mining, are effective in the prediction of a river flood[Chaitanya Damle and Ali Yalcin, 2007].Cluster based neural network model is effective in capturing nonlinear relationships among many hydrological processes.[Kamban parasuraman et al.,2007]. Data mining functionalities and the kinds of patterns they can discover are:

- Concept/class description
- Association analysis
- Classification and prediction
- Cluster analysis
- Outlier analysis

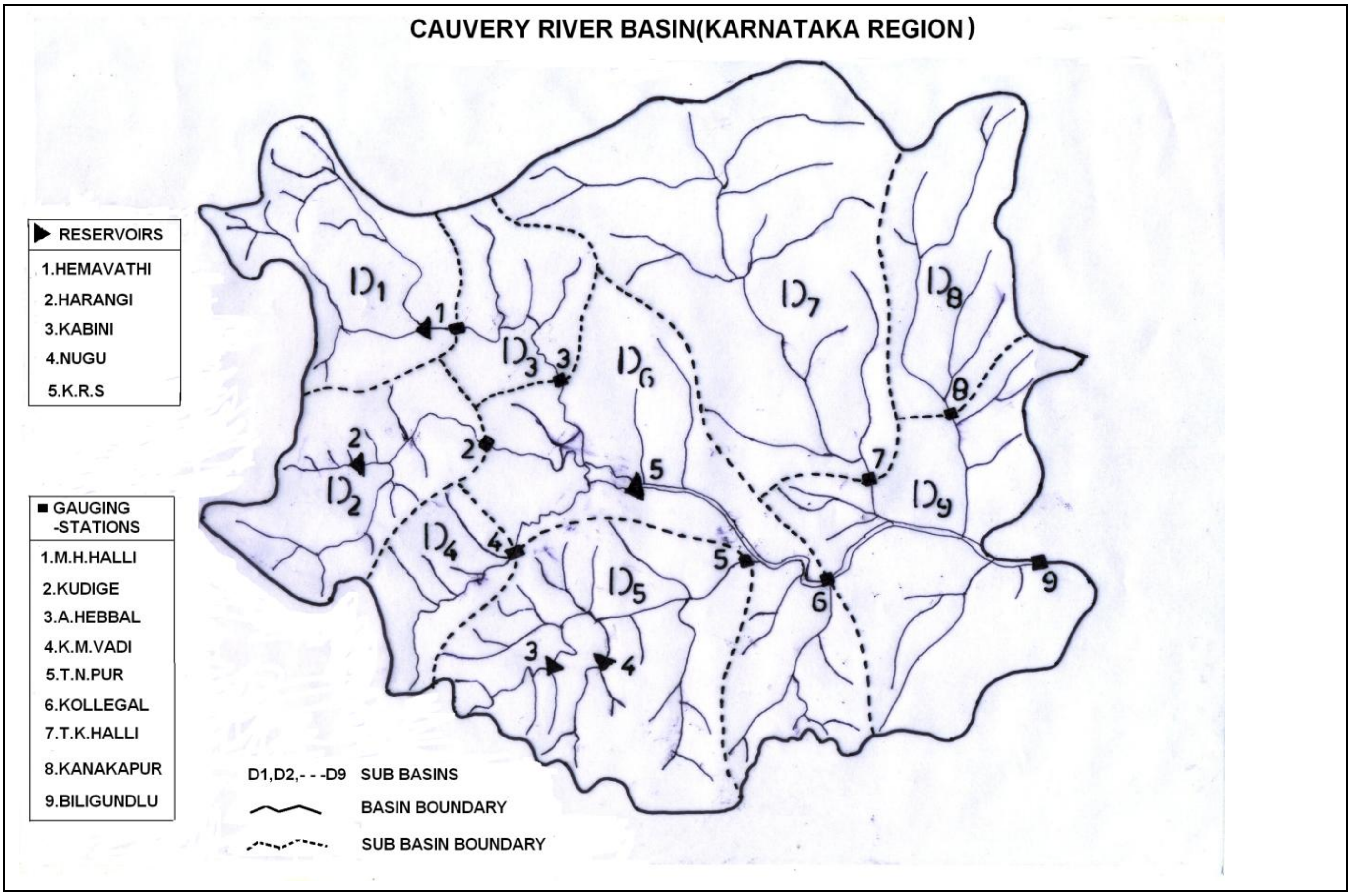

Fig 1: Cauvery Basin Map Showing Sub Basins

2.1 ASSOCIATION RULE MINING

Association rule mining searches for interesting relationships among attributes in a dataset. Association rules are similar to classification rules except that they can predict any attribute not just the class, and this allows them to predict combination of attributes too. Association rules are not intended to be together as a set. Different association rules express different regularities that underlie the dataset, and they generally predict different things.

Because so many interesting association rules can be derived from even a tiny dataset, interest is restricted to those that apply to a reasonably large number of instances and have a reasonably high accuracy on the instances to which they apply to. The coverage of an association rule is the number of instances for which it predicts correctly. [Zhou,H. et.al., 2008] This is often called its support. Its accuracy often called confidence is the number of instances it predicts correctly expressed as a proportion of all instances to which is applies.

The rule Rainfall (T,”LESS”) → Discharge (T,”LESS”) means if rainfall, T, is less then discharge, T is also less.

The accuracy is the proportion of the days when rainfall is less than the mean rainfall also has discharge less than the mean discharge, expressed in percentage or fraction. It is usual to specify minimum support (coverage) and the confidence (accuracy) values and to seek only those rules whose support and confidence are at least equal to these specified minima. Rules that satisfy both minimum support threshold and minimum confidence threshold are called strong. Generally support and confidence values are expressed between 0% to 100% rather than 0 to 1.0.

There are two methods for mining the simplest form of association rules-single dimensional, single level, Boolean association rules. [Harms,S.K.and Deogun,J.S. 2004] One is a basic algorithm for finding frequent item sets and another one is the frequent pattern growth method which adopts a divide and conquers strategy. Apriori algorithm [Witten,I.H. et al., 2008]. for mining frequent item sets for Boolean association rules is used in the present study. The algorithm employs an iterative approach known as level-wise approach where k item sets are used to explore (k+1) item sets. In Direct marketing Association rules are useful to summarize customer groups and to build a model for profit prediction [Wang K S Zhou, et al., 2005]

3. STUDY AREA AND DATA

In the present study, monthly cloud cover, potential evapotranspiration (PET), temperature, vapour pressure, rainfall and stream flow data (1980-2002) of eight sub basins of Cauvery river in the Karnataka(state) region, India are used. The river Cauvery originates at Talakaveri in Coorg district of Karnataka in Bhramagiri range of hills in the Western Ghats at an elevation of 1314.1m and drains total of 81,155sq.kms area of which 34273sq.kms is in Karnataka state, 43856sq.kms in Tamilnadu state , 2866sq.kms in Kerala state and 160sq.kms in the union territory Pondicherry.

The Cauvery basin is fan-shaped in Karnataka and leaf-shaped in Tamilnadu. The runoff does not drain off quickly because of its shape and therefore, no fast rising floods occur in the basin. The basin receives rainfall mainly from the S-W monsoon and partially from N-E monsoon in Karnataka. The basins in Tamilnadu receive good flows from the north-east monsoon. The sub basins considered for the present study with tributaries and gauge sites are provided in table 1.

4. APPLICATION AND RESULTS

The eight sub basins of Cauvery river as shown in Fig 1. Twenty three years(1980-2002) monthly data of cloud cover, PET, RCET, temperature, vapor pressure and rainfall for each sub basin are considered for the present study.

Various combinations of input data were tried and for each of them association rules were generated using weka environment. Fig.2 shows the sample output of the association rules(obtained) for sub basins D1 to D9(except D3 due to lack of available data) for peak flow months of 23 years(1980-2002) with minimum support=0.25 and minimum confidence=0.9

The generated association rules indicate that sub basins D2,D4 and D6 are strongly associated with peak flows in the same months.

The association rules also show that there is some association between sub basins D2, D4 and D5 and also between D2, D4 and D9 with respect to peak flow months.

Sub basin 'D1' is not associated with any other sub basin.

Table 1. Sub basins with gauge sites and tributaries

L. NO	Sub basin index	Gauge site	Tributary
1	D1	M H Halli	Hemavathi
2	D2	Kudige	Harangi
3	D4	K M Vadi	Lakshmana thirtha
4	D5	T N Pura	Katrini
5	D6	Kollegal	Cauvery
6	D7	T K Halli	Cauvery
7	D8	Kanakapur	Cauvery
8	D9	Biligundlu	Cauvery

Similarly the association rules are also generated for sub basins D1 to D9 (except D3) for low flow months and for mean monthly flows for MORE or LESS values corresponding to respective monsoon/non monsoon mean of 23 years for the months January to December indicate strong association between sub basins D2, D4 and D6.

```
=== Run information ===

Scheme:
weka.associations.Apriori -N 10 -T 0 -
C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c
-1
Relation:     sub basins with max
flows
Instances:    26
Attributes:   9
              YEAR/S.BASIN
               D1
               D2
               D4
               D5
               D6
               D7
               D8
               D9
=== Associator model (full training
set) ===


Apriori
=======

Minimum support: 0.25 (6 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15
```

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 15

Size of set of large itemsets L(3): 8

Size of set of large itemsets L(4): 1

Best rules found:

 1.D4=AUGUST 9 ==>D5= SEPTEMBER 9
conf:(1)
 2.D2=AUGUST D4= AUGUST 7 ==>  D5=
SEPTEMBER 7    conf:(1)
 3.  D5= SEPTEMBER  D9=AUGUST 7 ==>
D2=AUGUST 7    conf:(1)
 4.  D2=AUGUST  D9=AUGUST 7 ==>  D5=
SEPTEMBER 7    conf:(1)
 5.  D6= JULY 6 ==>  D2= JULY 6
conf:(1)
 6.  D6= JULY 6 ==> D4= JULY 6
conf:(1)
 7. D4= AUGUST  D9=AUGUST 6 ==>
D2=AUGUST 6    conf:(1)
 8. D4= JULY  D6=JULY 6==> D2= JULY 6
conf:(1)
 9.  D2= JULY D6= JULY 6==>D4=JULY   6
conf:(1)
10.  D6= JULY 6 ==> D2=JULY D4= JULY 6
conf:(1)
```

Fig.2 Association rules for sub basins D1 to D9 for peak flow months

Fig.3 shows a graph of number of common association rules v/s pair of sub basins for mean flow values. The graph illustrates the similarity between sub basins D2 and D4 with five common Association rules and between D1 and D5 with four common association rules.

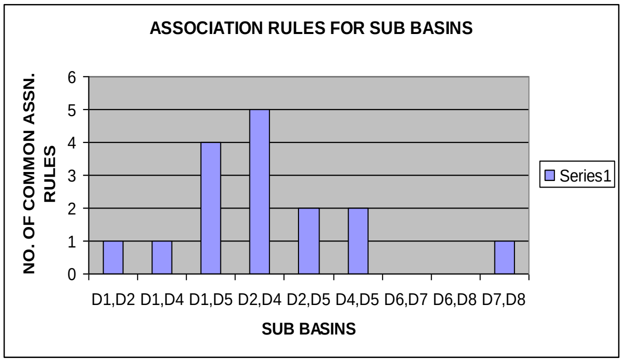

Fig.3 No. of common Association Rules v/s Sub basins.

5. CONCLUSIONS

Previous studies considered that sub basins of a river are homogeneous for hydrological analysis. This paper demonstrates the application of association rules of data mining for characteristic study of sub basins of a river. The association rules generated indicates the relationships between climatic factors and rainfall and discharge for each sub basin. The developed rules indicate the sub basins D2 and D4 and also D1 and D5 are hydrologically homogeneous. The obtained association rules show that the effect of discharge of sub basin D1 on the d/s sub basin D6 is negligible. This can be substantiated by the presence of reservoir just u/s of gauging station in sub basin D1. Also from the association rules it can be concluded that the discharge of d/s sub basin (D9) depends on discharge of sub basins D6 and D7 and not on discharge of sub basin D8.

6. REFERENCES

[1] Han,J. and Kamber,M. 2006. Data Mining: Concepts and Techniques,ISBN 1-55860-489-8,550,Morgan Kaufmann Publishers.

[2] Piatetsky-Shapiro,G. 1991. Discovery analysis and presentation of strong rules, Knowledge Discovery in Databases, AAAI/MIT Press.pp.229-248.

[3] Thandaveswara B.S. et al., 2000,Clarification of River basins using Artificial Neural Networks, ASCE Journal of Hydrologic Egg ,pp.290-298.

[4] Nagesh kumar,D. and Dhanya,C.T. 2009. Data Mining and its Applications for Modeling Rainfall Extremes, ISH Journal of Hydraulic Engineering,Vol.15,No.SP.1,pp.25-51.

[5] Chaitanya Damle and Ali Yalcin, 2006, Flood prediction using Time Series Data Mining, Journal of Hydrology, Vol.333,pp.305-316.

[6] Kamban parasuraman et al., 2007, Cluster Based Hydrologic prediction using Genetic Algorithm Trained Neural Networks, ASCE Journal of Hydrologic Engineering., vol.12, issue 1 ,pp.52-62.

[7] Zhou,H. etal. 2008. Time related Association rules Mining with attributes Accumulation Mechanism and its application to Traffic Prediction, Journal of Advanced Computational Intelligence and Intelligent Informatics,Vol.12,NO.5,pp.467-478.

[8] Harms,S.K.and Deogun,J.S. 2004. Sequential association rule mining with time lags, Journal of Intelligent Information Systems,Vol.22,No.1,pp.7-22.

[9] Witten,I.H. and Frank,E. 2008. Data Mining: Practical Machine Learning Tools and Techniques,ISBN 978-81-312-0050-6,Morgan Kaufmann Publishers.

[10] Wang K.S. Zhou, et al., 2005 Mining customer value from association rules to direct marketing,Data Mining and Knowledge Discovery,Vol.11,pp. 57-79.