Indian Journal of Science and Technology, *Vol 9(10), DOI: 10.17485/ijst/2016/v9i10/82155, March 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Water Quality Assessment using Association Rule Mining for River Narmada

Sanjeev Gour[1*], Shailesh Jaloree[2] and Mamta Gour[3]

[1]Department of Computer Science, L. B. S. College of Professional Studies, Harda - 461331, Madhya Pradesh, India; sunj129@gmail.com
[2]Department of Applied Mathematics, S. A. T. I. Engineering College, Vidisha - 464001, Madhya Pradesh, India
[3]Department of Chemistry, Government P. G. College, Harda - 461331, Madhya Pradesh, India

Abstract

Objective: The objective of the this study is to analyze the various water quality parameters of the Narmada River and to find hidden relationship between them so that it suggest some decision plans or policies to predict or classify the water quality. **Methods:** In this study we find an approach to water quality management through Association or correlation studies between various water quality parameters. The Data Mining Technique called Association Rule Mining (Apriori Algorithm) is used to find and extract some rules or relationship between various water quality parameters for Narmada River at Harda and Hoshangabad districts of Madhy Pradesh. **Findings:** We have found some interesting and useful correlation between different water quality parameters and also we measure the performance of algorithm by the confidence and lift value factors. **Application:** This research present a model with actual data both for spatial and temporal patters and benefits of employing data mining techniques towards the improvement of water quality management plans. These results conclude that there is urgent need of strict regulatory monitoring for water quality maintenance in the river system at Hoshangabad District.

Keywords: Apriori Algorithm, Association Rule Mining

1. Introduction

Water is a fundamental need of human life and therefore is the essential for human survival and development. The need of fresh water resources, their utilization and conservation have very important consideration during the present time. The river Narmada is the biggest west flowing river of the state and fifth largest west flowing river of India. There are huge changes in river water quality in during last 10 years. This may be due to great involvement of human activities and industries waste. State Government has needed to be made effective monitoring plan for fresh and clean river water. Many recent research activities have implemented in this area. Some of research studies[5] have been conducted in a scientific way to help the management team to sort out the hidden information from the data which are generated during the everyday. Anonymization technique using sub clustering is specified which achieves hidden relationship between water quality parameters with maximum privacy[6] and also using by Apriori Algorithm[7].

2. Fundamental of Data Mining

The method to extract information from repository of data stored in a database is known as data mining[1]. Nowadays Information Technology (IT) and its application areas includes a huge available database, together with the data mining methods are used to retrieve and interpret important facts that is available in the databases, and also to extract the essential information and their association to generate information or knowledge which is useful for decision making plans.

3. Association Rule Mining: Concept

The process to find association rules by showing conditions of attribute-value which frequently occur together

*Author for correspondence

in a given database is called association mining. The rules have a form: If item A is part of an event, then X% of the time item B is also part of the event. The rules are written as A ⇒ B, where A is called the antecedent and B is called the consequent. The associations rule $A \Rightarrow B$ is interpreted as database tuples that satisfy the condition in A are also likely to satisfy the condition in B.

As State by Witten[2], an item in rules is an attribute-value pair. The confidence is the conditional probability of B given A, P (B/A). A rule is "interesting" if the conditional probability P (B/A) is significantly different than P (B). Confidence of the rule measures the rule's accuracy.

Association algorithms discovering these innovative rules by processing the equivalent of sorting the data while counting occurrences such that they can calculate confidence and support. The efficiency during the extracting association's rules is one of the differentiators among algorithms. As explored by[3], association rules are considered interesting if they follow both a minimum support threshold and a minimum confidence threshold and hence considered strong rules.

As stated earlier, an association rule is a Statement $X \Rightarrow Y$, where X and Y are sets of items. Given a database of transactions-where each transaction $T \in D$ is a set of items, statement $X \Rightarrow Y$ shows that whenever a transaction T includes X then T probably includes Y also. The probability is the percentage of transactions containing Y in addition to X with respect to the total number of transactions containing X. This concept of mining association rules emitted from the popular analysis of market basket data[4].

4. Preparation of Data and Data Mining Software

The method of data filtering and cleaning depends on the data mining algorithm specifically and on the data Mining software which is used for the problem domain. For data mining problem areas, WEKA is an open source collection of machine learning algorithms, which also contains Association mining capabilities. WEKA is multi-functional data mining software. Data preparation, classification, association, clustering and visualizing input and output are the main functions contained in the software. In WEKA, Apriori is the only association rule algorithm. In this experiment with the help of WEKA preprocessing tools, we prepared and transform data from numeric to nominal. (Figure 1). The consistency of individual attribute

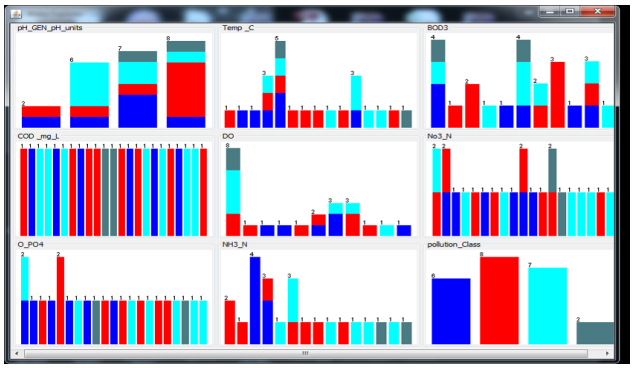

Figure 1. Preprocessing mechanism in WEKA.

values, distribution of missing values, types, quantity etc. are checked after receiving the target dataset.

5. Apriori Algorithm

This algorithm is one of the best and base association rule mining algorithms for most parallel algorithms. This algorithm uses a bottom-up search and enumerates all frequent item sets completely. It is completely based on data passes. It finds frequent "item sets", subsets of items with a transaction T, by processing as many data passes as given by the user, or until there are no extra frequent item sets to be identified. Thus, the procedure starts by searching all transactions T in the database D and calculating the frequent items. Later, a set of frequent item sets (i.e., Set-2) is formed from the frequent items. For the next pass, the frequent item sets (Set-2) are retained, and the procedure repeats until all frequent item sets have been enumerated. It might be quite costly because the produced rules normally are in very large and in contrast the percentages of relevant and useful rules are in very small fraction. In this way we need to mainly focus on supporting the user when browsing the rule set and the development of further useful quality measures on the rules[4].

6. Experiment Data Set

In this experimental study, we have used Water Quality Data of District Harda and Hoshangabad of M. P., (India) from the analysis period from 1990 to 2010 of River Narmada.

7. Experimental Setup

Firstly, we have explored features of WEKA and its function by importing dataset samples and then with the software

we tried to open the learning datasets. After this we tried to find some interesting association rules based on a particular learning dataset. Our experiment is limited upto a pre-given dataset: (Water Quality Data of Handia (District Harda) and Hoshangabad, M. P., from the analysis period 1990 to 2010 of River Maa Narmada). WEKA Explorer runs environment for Association and Parameters setting during experiments with WEKA are shown in Figure 3 and 4.

8. Selected Parameters

There are many parameters which may affect the river water quality. In this experimental study, we have taken into consideration nine parameters for the investigations after data pre-processing. Attribute view after Pre-processing in WEKA can be seen in Figure 2.

9. About Class Attribute

On the basis of standards pollutant index (Table 2.), surface water quality can be categorized into four classes;

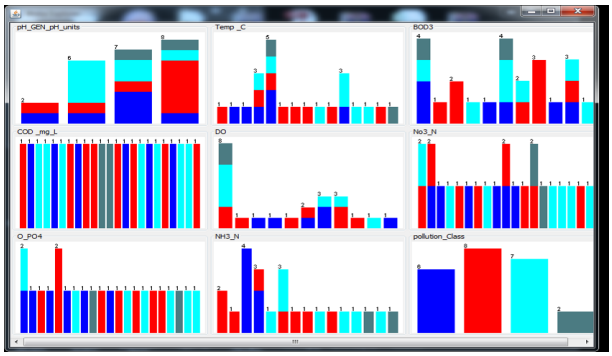

Figure 2. Attribute view after preprocessing in WEKA.

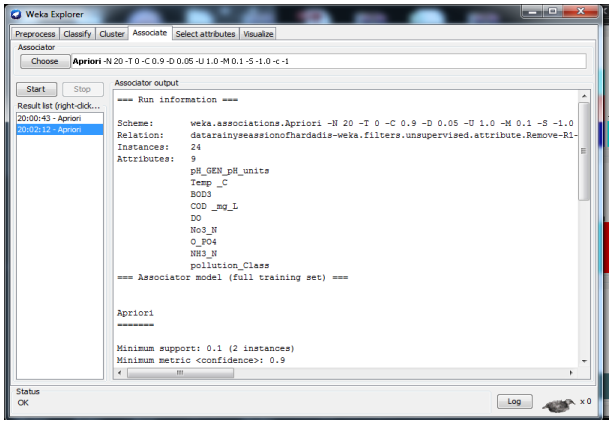

Figure 3. WEKA explorer run environment for association.

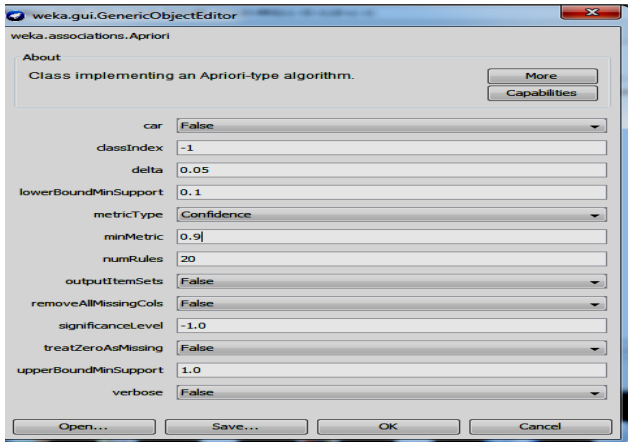

Figure 4. Parameters setting during experiments with WEKA

Table 1. List of selected parameters

Sl. No.	Attribute	Abbreviation
1	PH	PH value
2	DO	Dissolve oxygen
3	BOD	Biochemical Oxygen Demand
4	No3_N	Nitrate Nitrogen
5	NH3_N	Ammonia Nitrogen
6	TEMP	Temperature
7	COD	Chemical oxygen demand
8	PO4	phosphate
9	Class (Polluted class)	A,B,C and D

Table 2. Pollutants index 2 class value for

Pollutants index	Class			
	A	B	C	D
PH(Mg/l)	<5	5-9	5-9	5-9
DO(mg/l)	>6	6	4	2
BOD (mg/l)	<1.5	1.5	2	4
(NO3-N(mg/l)	<5	5	5	5
NH3N(mg/l)	<0.5	0.5	0.5	0.5

Class A: Perfectly clean fresh surface water use for consumption and is not necessary to pass it through any water treatment process.

Class B: Excellently clean fresh surface water use for consumption which requires ordinary water treatment process.

Class C: Moderately clean fresh surface water use for consumption, but recommended to pass through an ordinary treatment process before use;

Class D: Finely clean fresh surface water use for consumption, but strictly requires special water treatment process before use.

10. Experimental Results

Total Instances: 24, Consider Parameters: 9, Algorithm used: Apriori

=== Associator model (full learning set) ===

Minimum support: 0.1 (2 instances), minimum <confidence>: 0.9, Number of cycles performed: 18

Generated sets of large item sets:

Size of set of large item sets L (1): 31, Size of set of large item sets L (2): 41

Size of set of large item sets L (3): 12, Size of set of large item sets L (4): 1

Some of Basic Interesting rules found by Experiment are:

11. Conclusion

We extracted some of the relation between various water quality parameters using association rules. Table 3 is the summary of the result of this study.

We have found that if the concentration of NH3-N decreases by some level, then water quality goes poor by one level. It is also showed that for a constant value of pH, BOD is strongly reverse to DO. It is clear by the rules extracted by mining that the amount of DO and water pollution has a reverse relation that is decrease of DO concentration increases the water pollution. We have also found the rule that show the relationship between BOD and DO which indicates that if amount of DO goes less, then for the same value of BOD the quality of water goes poor at one level.

We have concluded that water quality improved whenever the quantity of BOD decreases for the same value of pH. In other words if the water contains more BOD concentration, then the surface water quality decrease proportionally.

From this result we have found that BOD concentration is more at Hoshangabad District as compared to Harda District that is why the surface water quality is poor in Hoshangabad at one level of class. We have also found the relation between the parameters NH3_N and O_PO4 that is if NH3_N decreases then O_PO4 causes water quality get poor by one level.

Table 3. Association rules found in experiment

Sl. No.	Rule	Confidence	Lift value
1	NH3_N=0.02 4 ⇒ pollution_Class=A 4	1	4
2	BOD3=1.4 3 ⇒ pollution_Class=B 3	1	3.43
3	pH_GEN_pH_units=8.2 BOD3=1.2 3 ⇒ DO =6.1 3	1	3
4	pH_GEN_pH_units=8.2 BOD3=1.2 3 ⇒ DO =6.1 3	1	3
5	DO=6.9 2 ⇒ Temp _C=27.3 2	1	4.8
6	BOD3=0.9 2 ⇒ DO=7.4 2	1	8
7	BOD3=0.9 2 ⇒ pollution_Class=B 2	1	3
8	pollution_Class=D 2 ⇒ DO=6.1 2	1	3
9	No3_N=0.16 2 ⇒ DO =7.4 2	1	8
10	NH3_N=0.01 2 ⇒ O_PO4=0.021 2	1	12
11	O_PO4=0.021 2 ⇒ NH3_N=0.01 2	1	12
12	O_PO4=0.021 2 ⇒ pollution_Class=B 2	1	3
13	NH3_N=0.01 2 ⇒ pollution_Class=B 2	1	3
14	NH3_N=0.05 pollution_Class=C 2 ⇒ pH_GEN_pH_units=8.1 2	1	4
15	pH_GEN_pH_units=8.1 NH3_N=0.05 2 ⇒ pollution_Class=C 2	1	3.43
16	BOD3=0.6 pollution_Class=A 2 ⇒ pH_GEN_pH_units=8.2 2	1	3.43
17	pH_GEN_pH_units=8.2 BOD3=0.6 2 ⇒ pollution_Class=A 2)	1	4
18	BOD3=1.2 pollution_Class=C 2 ⇒ pH_GEN_pH_units=8.2 2		
19	pH_GEN_pH_units=8.2 pollution_Class=C 2 ⇒ BOD3=1.2 2	1	6
20	pH_GEN_pH_units=8.2 pollution_Class=C 2 ⇒ DO =6.1 2	1	3

Our Experiments confirms the Dependence between inclusions of industrial centers and DO (Dissolve Oxygen). Sewage waste water of industries in the river Narmada at Hoshangabad has significant pollutants of these water quality parameters and therefore concentration of industrial waste chemicals increases these types of pollutants in the Narmada River. It is conclude that the concentration of DO and water pollution have a reverse relation, that is decrease of DO increases the water pollution.

Many other important water quality parameters should be considered for extracting more hidden relationship and their effect on the River should also be taken into consideration for the research studies in future.

12. References

1. Miller HJ, Han J. Discovering geographic knowledge in rich environment. SIGKKD Exploration; 2001. p. 105–8.
2. Witten H, Frank E. Data mining: Practical machine learning tools and techniques. 3rd ed. Morgan Kaufmann Publishers; 2011. ISBN: 978-0-12-374856-0.
3. Han J, Kamber M. Data mining: Concepts and techniques. 2nd ed. Morgan Kaufmann Publishers, University of Illinois at Urbana-Champaign; 2001.
4. Hipp J, Guntzer U, Nakhaeizadeh G. Algorithms for association rule mining: A general Survey and comparison. ACM SIGKDD Explorations Newsletter. 2000; 2(1):58–64.
5. Shankar R, Sundararajan M. Manufacturing quality improvement with data mining outlier approach against conventional quality measurements. Indian Journal of Science and Technology. 2015 Jul; 8(15). DOI: 10.17485/ijst/2015/v8i15/73109.
6. Rajalakshmi V, Mala GSA. Anonymization by data relocation using sub-clustering for privacy preserving data mining. Indian Journal of Science and Technology. 2014 Jan; 7(7). DOI: 10.17485/ijst/2014/v7i7/44454.
7. Jafarzadeh H, Torkashvand RR, Asgari C, Amiry A. Provide a new approach for mining fuzzy association rules using apriori algorithm. Indian Journal of Science and Technology. 2015 Apr; 8(S7). DOI: 10.17485/ijst/2015/v8iS7/71227.