Binary Journal of Data Mining & Networking 4 (2014) 25-28

http://www.arjournals.org/index.php/bjdmn/index

Research Article

ISSN: 2229 -7170

Decision tree approach to build a model for water quality

Shailesh Jaloree[1], Anil Rajput[2], Sanjeev Gour[3]

*Corresponding author:

Shailesh Jaloree

[1]S.A.T.I. Engineering College.Vidisha (M.P.) India
[2]Govt.P.G.College, Sehore M.P.India
[3]Barkatullah University, Bhopal, India

Abstract

This paper presents a Classification data model using decision tree for the purpose of analyzing water quality data of MAA Narmada River at Harda district. The data model was implemented in WEKA software. Classification using decision tree was applied to classify /predict the pollutant class of water. It is observed in the analysis that the Nitrogen (NH3_N ,NO3_N), pH ,Temp _C, BOD, COD, other parameter relevant to water processes play an important role to assess the quality of river water. In this experiment we have used five attribute of water quality data which can affect accuracy of water.

Keywords : Data Mining, classification model, Decision tree, Weka Tool.

Introduction

The rivers of India play an important role in the lives of the Indian people. The river systems provide irrigation, potable water, cheap transportation, electricity, as well as provide livelihoods for a large number of people all over the country.

Water pollution too adds to the existent problems of local and regional water scarcity by making large amounts of water unfit for consumption. With increasing agricultural, industrial and domestic needs, there is also a growing competition for clean water supplies. Surface water systems are facing significant disturbances via reclamation, alteration and pollution due to the increasing pressures from urban expansion and urban land use change. In one way or another, many water related problems (e.g. flooding or drought disasters, and serious water pollution) are the outcome of disordered or ill-conceived land use development [1]. However, the linkage between land and water resource management in the urban area has long been ignored. Under the circumstances of extensive urbanization in India, many cities are experiencing the serious impacts of rapid urban land use expansion on surface water systems.

Rapid growth in industrialization to support the country's growing population and economy has polluted our rivers like never before. Studies show that domestic and industrial sewage, agricultural wastes have polluted almost all of Indian rivers. Most of these rivers have turned into sewage carrying drains. This poses a serious health problem as millions of people continue to depend on this polluted water from the rivers. With expansion in monitoring of water systems data-driven techniques are becoming more interesting and useful. In particular, data mining is a clear approach to investigate, as it can deal with a high degree of complexity within the given data[2]. Data mining is the search for valuable information in large data sets, trying to discover patterns in the data. Data mining techniques can be used to classify data records and to allow for the creation of new hypotheses about the system behavior. The present study is an effort to assess and model the water quality of the MAA NARMADA, west flowing river of the state of Madhya Pradesh.

Data Mining Concept

Data mining is an approach for information extraction from huge amount of data stored in a database [3]. Recent trends in information technology (IT) and its growing application areas in addition to increase of available databases, along with the data mining are being used to extract and interpret information available in the databases, and explore the necessary information and their relationships to produce useful information/knowledge for decision making.

Definition Translating Data mining word by word means the mining or digging in data with the purpose of finding information or respectively knowledge. Coming to the more abstract and very well known definition of Frawley, Data mining is defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [4].

Decision Tree

Decision tree is a machine learning technique that allows us to estimate a quantitative target variable (for example, profit, loss or loan amount) or classify observation into one category of a categorical target variable (for example, good/bad credit customer; churn or do not churn) by repeatedly dividing observations into mutually exclusive groups. The algorithm commonly used to construct decision tree is known as recursive partitioning and the common algorithms are CHAID (Chi-square Automatic Interaction Detection), CART (Classification & Regression Tree) and C5.0. Decision trees represent a supervised approach to classification.

A decision tree is a decision support system that uses a tree-like graph decisions and their possible after-effect, including chance event results, resource costs, and utility. A Decision Tree, or a classification tree, is used to learn a classification function which concludes the value of a dependent attribute (variable) given the

values of the independent (input) attributes (variables). This verifies a problem known as supervised classification because the dependent attribute and the counting of classes (values) are given [5].

Decision trees are the most powerful approaches in knowledge discovery and data mining. It includes the technology of research large and complex bulk of data in order to discover useful patterns. This idea is very important because it enables modelling and knowledge extraction from the bulk of data available. All theoreticians and specialist are continually searching for techniques to make the process more efficient, cost-effective and accurate. Decision trees are highly effective tools in many areas such as data and text mining, information extraction, machine learning, and pattern reorganization.

Decision tree offers many benefits [6] to data mining, some are as follows:-

- It is easy to understand by the end user.
- It can handle a variety of input data: Nominal, Numeric and Textual
- Able to process erroneous datasets or missing values
- High performance with small number of efforts
- This can be implemented data mining packages over a variety of platforms

A tree includes: - A root node, leaf nodes that represent any classes, internal nodes that represent test conditions (applied on attributes) .

Weka Software

The WEKA software was developed in the University of New Zealand. A number of data mining methods are implemented in the WEKA software. Some of them are based on decision trees like the J48 decision tree, some are rule-based like ZeroR and decision tables, and some of them are based on probability and regression, like the Naïve bayes algorithm. Weka uses the J48 algorithm, which is Weka's implementation of C4.5 Decision tree algorithm. J48 is actually a slight improved to and the latest version of C4.5. It was the last public version of this family of algorithms before the commercial implementation C5.0 was released.

Experimental Datasets

In this research we have used Water Quality Data of Narmada river of Harda District M.P from 1990 to 2010. The dataset after preprocessing is given in table 1.

Experimental Setup

In the Experiment whole datasets was training as a training set for developing a model. J48 decision tree classifier were used in this study.J48 is an open source Java implementation of the C4.5Algorithm in the Weka data mining tool. C4.5 is a program that creates a decision tree based on a set of labeled input data. This algorithm was developed by Ross Quinlan. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

Data Preparation

The process of data cleaning and preparation is highly dependent on the specific data mining algorithm and software chosen for the data mining task. Here i attempted to prepare the data according to the requirements of the selected data mining software.

Pre Processing of the Data

A data set collected is not directly suitable for induction (knowledge acquisition), it comprises in most cases noise, missing values, and inconsistent data set is too large, and so on. Therefore, we need to minimize the noise in data, choose a strategy for handling missing (unknown) attribute values, use any suitable method for selecting and ordering attributes (features) according to their informatively (so-called attribute mining), discredited/fuzzily numerical (continuous) attributes, validating part of training data to be used for creating model and eventually process continuous classes.[7]

Year	PH_gen	DO	BOD	No3_N	NH3_N	Class
1990	8.2	8.1	0.8	0.39	0.08	I
1991	7.2	8.3	0.9	0.52	0.02	II
1992	7.9	0.9	1.4	0.41	0.03	I
1993	8.2	7.8	1.8	0.55	0.02	III
1994	8.3	7.9	1.3	0.56	0.01	IV
1995	8.3	8.1	1.9	0.58	0.03	III
1996	8.9	8.9	1.3	1.44	0.03	II
1997	8	7.5	0.9	0.47	0.02	I
1998	8.3	8.3	0.5	0.51	0.12	IV
1999	8.1	7.9	0.9	0.25	0.05	I
2000	8.4	8.2	0.6	0.67	0.03	II
2001	8.3	8.1	1.5	0.44	0.03	IV
2002	8.3	7.8	1	0.28	0.01	IV
2003	8.4	7.7	1.3	0.51	0.02	II
2004	8.1	8.1	0.6	0.52	0.03	II
2005	8.2	7.5	1.1	0.43	0.02	IV
2006	8.2	8.1	1.3	0.06	0.01	IV
2007	7.9	7.6	0.8	0.49	0.02	II
2008	8.3	7.3	1.5	0.35	0.03	IV
2009	8.3	7.2	1.2	0.51	0.12	IV
2010	8.4	7.4	1.2	0.31	0.13	IV
2011	8.1	7.3	1.3	0.85	0.02	III
2012	8.0	8.2	1.3	0.56	0.12	III

Table 1-Water quality dataset after pre-processing

Water quality parameters which is used for creating Model

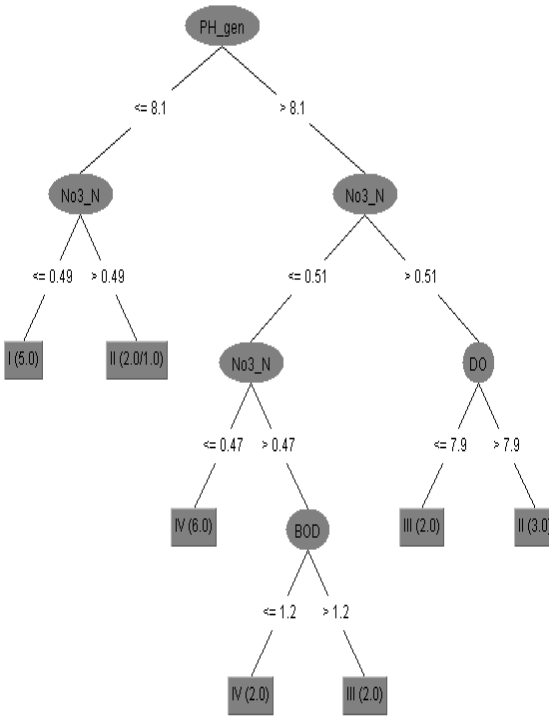
Many parameters can influence the surface water quality. Five parameters are selected for the investigations. The surface water quality can be classified as in following Table 2.

S.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	Class (Polluted class)	I,II,III,IV

Table 2-The surface water quality parameter

Description Of Class Attribute

Generally, surface water quality can be divided into five classes; class I , extra clean fresh surface water resources use for conservation that are not necessary to pass through water treatment processes and require only ordinary processes for pathogenic destruction and ecosystem conservation where basic organisms can breed naturally; class II, very clean fresh surface water resources use for consumption that require ordinary water treatment processes before use by aquatic organisms in



conservation, fisheries and recreation; class III, medium clean fresh surface water resources use for consumption, but are passed through an ordinary treatment process before use; class IV, fairly clean fresh surface water resources use for consumption, but requires special water treatment processes before use . Class attribute have created according to following table 3.

Table 3- Class value for pollutants index

Water Quality Parameter	Class I	II	III	IV
PH (mg/l)	<5	5-9	5-9	>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
NH3_N(mg/l)	<0.5	0.5	0.5	0.5

Figure 1- WEKA generated tree

We have found a classification model. If...Then rules can be extracting from weka generated tree. Some interesting rules of experiment are as follows:

Rule 1 : IF PH_gen>8.1 and NO3N>0.51 and DO>7.9
THEN Class- II (75%)

Rule 2: IF PH_gen>8.1 and (NO3N<=0.51 but NO3N>0.47) and BOD>1.2
THEN Class- III (40%)

Rule 3: IF PH_gen>8.1 and (NO3N<=0.51 but NO3N>0.47) and BOD<=1.2
THEN Class -IV (25%)

Rule 4: IF PH_gen>8.1 and (NO3N<=0.51 but NO3N<=0.47)
THEN Class -IV (75%).

Rule 5: IF PH_gen <=8.1 and NO3N<=0.4.9
THEN Class -I.

Rule 6 : IF PH_gen <=8.1 and NO3N>0.4.9
THEN Class –II .

Rule 7 : IF PH_gen>8.1 and NO3N>0.51 and DO<=7.9
THEN Class- III .

Number of Leaves: 7

Size of the tree: 13

Time taken to build model: 0.08 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances	21	95.4545 %
Incorrectly Classified Instances	1	4.5455 %

Table 4-Accuracy of classification of instances during experiment

Conclusions

In this result we have implemented water quality data with decision tree technique. We have used J48 classifier in WEKA data mining

tools. Experiment we have used five attribute of water quality data which can affect accuracy of water. We have found a model. .we have found that the correctly classification of instances is 95.4545% and incorrectly classification is 4.5455% *(See Table 4).* If...then rules can be extracting from weka generated tree. Some interesting rules of experiment are given above. we have some extracted knowledge about quality of water are: If pH value goes to less than or equal to 8.1 unit and the amount of NO3 N increase then the quality of surface water decrease at one class level. That means amount of NO3 N increases causes increase pollution in surface water. If pH value goes to greater than 8.1 and the amount of NO3-N lies between 0.47 and 0.51 then decrease of amount in BOD also Decrease the quality of water at one level class. While as increase the value in NO3-N and DO in surface water, the quality of water also improved at one level of class.

References

[1]. Dehalwar kavita and Singh Jagdish ,2012-"Water Resources Management and Water Quality, case of Bhopal" , International Conference on Chemical, Ecology and Environmental Sciences (ICEES'2012) march 17-18, 2012 Bangkok.

[2]. Fayyad U.M., Piatetsky-Shapiro G., Smyth P., (1996), From Data Mining to Knowledge.

[3]. Han, J. and M. Kamber, Data Mining: Concepts and Techniques Morgan Kaufmann, 2001.

[4]. Frawley William J., Gregory Piatetsky-Shapiro, and Christopher J. Matheus (1992) "Knowledge Discovery in Databases: An Overview.AI Magazine Volume 13 Number 3 .

[5]. Bhargava Neeraj et al ,2010- "Mining higher educational students data to analyze studentÊs admission in various discipline" . Binary Journal of Data Mining & Networking.

[6]. Bhargava Neeraj et al ,2013 –"Decision Tree Analysis on J48 Algorithm for Data Mining"-International journal of advance research in computer science and software engineering,vol-3,issue-6.

[7]. Manchanda sanjeev ,Dave Mayank and singh S. B. An Empirical Comparison Of Supervised Learning Processes "International Journal of Engineering, Volume (1) : Issue (1) .