

Application of Classification Techniques for Prediction of Water Quality of 17 Selected Indian Rivers



Harlieen Bindra, Rachna Jain, Gurvinder Singh and Bindu Garg

Abstract Objective: In this study, prediction using classification techniques are used to predict the water quality of the 17 selected rivers in the year 2011 using their water quality in 2008 to interpret whether the water quality has improved or deteriorated. Methods/Analysis: For this prediction, we have used data mining classification techniques using Waikato Environment for Knowledge Analysis (WEKA) API to the dataset of selected 17 Indian rivers. The data used for prediction was created from ambient water quality of Aquatic Resources in India in 2008 and 2011. Data is obtained from data portal which was published under National Data Sharing and Accessibility Policy (NDSAP) and the contributor was Ministry of Environment and Forests Central Pollution Control Board (CPCB). Findings: Out of the four techniques used, prediction of classes, i.e. excellent, good, average and fair is best done by Naive Bayes followed by J48, SMO and REPTree technique.

Keywords Prediction using classification techniques · Weka · Data mining Water quality · Indian rivers

1 Introduction

India is popularly referred to as the land of rivers since it has been blessed with several water bodies which not only enhance the beauty of the country but is also the source of livelihoods for a large number of people. They are the main sustainability source for people especially the farmers since the soil lands in proximity

H. Bindra (✉) · R. Jain · B. Garg
CSE Department, Bharati Vidyapeeth's College of Engineering, New Delhi, India
e-mail: harlieenbindra@gmail.com

R. Jain · G. Singh
CSE Department, Guru Tegh Bahadur Institute of Technology, New Delhi, India

to the rivers are nourished and fertile. For many holy reasons, these rivers are worshipped in India; specially “The Ganges” which is considered to be the holiest of all.

Indian Rivers not only nourish the flora and fauna but also attract tourist from all around the world and play an indispensable role in our economy. They are the witness of how the civilisation evolved but they are not only significant historically but also culturally and religiously. Even their inherent nature could not be altered by the dams. They still originate from the mountains and gush down the plains and valleys with the same force as several years ago. They nourish the plain with vitality and fertility.

But people day by day are forgetting the importance of rivers. The rivers now have fertilisers, pesticides and more different types of chemical products. A number of instances oil spills have disturbed the aquatic animals. The banks of the rivers are piled up with non-biodegradable wastes. But we need to understand that improving the unhygienic and dirty conditions of the rivers is not the sole responsibility of the government. We as the citizens of this nation should take special precautions and actions to improve the water quality of rivers. Even in western countries, the citizens themselves take measures to keep their rivers and river banks clean. We must strive to keep the best gift of nature clean and preserve its water quality.

For present study, 17 rivers were selected for prediction of water quality in 2011 using 2008 instances. The number of stations used to collect data for each river is mentioned in the parenthesis, which are

Beas (19), Satluj (20), Ganga (36), Yamuna (19), Brahmaputra (10), Dhansiri (7), Mahi (7), Narmada (6), Tapi (10), Mahanadi (14), Brahmani (11), Baitarni (5), Subarnarekha (6), Godavari (34), Krishna (22), Pennar (4), Cauvery (20). So, in total 250 instances were used for analysis.

In this paper, there are seven sections. Section 1 is the Introduction which is the current section. The Sect. 2 is Literature Review, the Sect. 3 is Materials and Methods, the Sect. 4 is Performance Comparison, the Sect. 5 is Result and Discussion, the Sect. 6 is Conclusion, seventh is Acknowledgement and the last section is References.

2 Literature Review

In paper [1], performance of CART, J48, REPTREE, Bayes Net and Naïve Bayes classification algorithms are compared by applying them to a dataset consisting of only 11 attributes, for predicting heart attacks. In the research work algorithms for prediction are applied using WEKA as it provides proficiency in analysing, discovering and predicting patterns. The results of the paper helped us in concluding that J48, CART and REPTREE shows the best results and there is not much difference in their performance factor. In paper [2], the author has compared the results of two decision trees ID3 and J48. The two techniques are applied to a dataset of students enrolling for MCA. The research work explains how tree based

classification algorithms ID3 and J48 works and are used to analyse the data. From the results it can be concluded that ID3 decision tree algorithm shows an accuracy of 69.69% as compared to that of J48 which is 67.67%. In paper [3] the two data mining algorithms which are used for producing the classification model are Naive Bayesian Classifier algorithm and Decision Tree algorithms. These algorithms are applied on preprocessed student dataset. Decision Tree algorithms has an accuracy of 93.33% over 71.67% of Naive Bayesian Classifier algorithm. Hence decision tree algorithm proves to be better than naïve Bayesian classifier. In paper [4], the author has explained about the heart diseases and symptoms of heart attack. The paper has talked about various models that are developed using different data mining techniques. In paper [5] has bestowed satisfactory modifications for calculation of the water quality index (WQI). To calculate the general water quality index nine parameters are required but sometimes a few parameters are missing or unavailable, in that case the modified formula given in this paper helps user to calculate WQI.

National River Conservation Plan [6] was initiated with the launching of Ganga Action Plan (GAP) in 1985. In 1995 GAP was expanded to cover other rivers of the country. At a sanctioned cost of Rs. 5779.41 crore, NRCP, excluding the GAP-I, GAP-II and National Ganga River Basin Authority (NGRBA) programme presently covers polluted stretches of 40 rivers in 121 towns spread over 19 States running head, it will be shortened. Your suggestion as to how to shorten it would be most welcome.

3 Materials and Methods

For present study, the data set was created using the data that referred to the ambient water quality of Aquatic Resources in India in 2008 and 2011 [7]. This Dataset is released under “National Data Sharing and Accessibility Policy (NDSAP)” and the contributor is “Ministry of Environment and Forests and Central Pollution Control Board”. The values of water quality parameters like Fecal Coliform, Temperature, Nitrate, Biochemical Oxygen Demand (B.O.D), pH, etc. were given in the data used. The data was published on the data portal on December 22, 2014 which was released under National Data Sharing and Accessibility Policy (NDSAP) [8] and the contributor was Ministry of Environment and Forests Central Pollution Control Board [9].

In classification [10] a set of objects is classified into a group so that objects in a group are more similar to each other.

3.1 *Classification Techniques Used*

3.1.1 Naïve Bayes [11]

Naïve Bayes Classifier is a part of probabilistic classifier based on application of Bayes' Theorem with strong independent presupposition/presumption between the features. This classifier algorithm presumes that in a given class attribute values are independent of other attributes values.

3.1.2 J48 [12]

J48 is the appendage of ID3. The features of J48 are decision tree pruning, keeping accounts for missing values, derivation of rules, etc. In WEKA, the implementation for JAVA open source algorithm C4.5 is done using J48. For tree pruning a number of options are provided by WEKA (data mining tool). Pruning could be employed as a mechanism for precisising if there is case of over fitting. The aim of this algorithm is to progressively generalise the decision tree till accurate and flexible tree is obtained. Continuous and discrete attributes can be handled by this algorithm.

3.1.3 SMO (Sequential Minimal Optimization) [13]

SMO stands for Sequential Minimal Optimization, John Platt invented this algorithm in 1998 at Microsoft Research. This algorithm is mainly used for solving quadratic programming problem which emerges at time of programming support vector machine. Nominal attributes are transformed into binary ones on implementation of the model. Also, by default it normalises all attribute. The worst case running time for this is $O(n^3)$.

3.1.4 REPTree [14]

REP stands for Reduce Error Pruning. This algorithm is based minimising error surfacing from variance and calculating the information gain using entropy. REP Tree generates various trees in reordered iterations and uses regression tree logic. It splits the missing values into pieces of corresponding instances.

3.2 *Software Used*

Eclipse [15] open source IDE (Integrated Development Environment) was used to compile the code wrote using Waikato Environment for Knowledge Analysis (WEKA) [16] (developed by the University of Waikato, New Zealand) API. It is commonly used for data mining works, as it has a number of machine learning algorithms. It has tools for preprocessing, classification, visualisation, etc. Eclipse is an IDE mostly used for computer programming in Java language but it also supports many other programming languages.

3.2.1 **General Algorithm**

1. Training dataset is loaded.
2. The class index is set to the last attribute.
3. Number of classes is fetched.
4. Class values in the training dataset is printed.
5. Class string value using the class index is fetched.
6. Creating and building the classifier.
7. The test dataset is loaded.
8. The class index is set to the last attribute.
9. Looping through the new dataset to make predictions.
10. Fetching class value for current instance.
11. Fetching class string value using the class index Class's int value is used.
12. Instance object of current instance is fetched.
13. Calling `classifyInstance`, which returns a double value for the class.
14. Use the double value to get string value of the predicted class.

3.2.2 **Code Used**

The following is the code used for Naïve Bayes Classifier. The classifier can accordingly be changed as per the requirement but the rest of the code will remain the same.

```
import weka.classifiers.bayes.NaiveBayes;  
import weka.core.Instance;
```

```

import weka.core.Instances;
import weka.core.converters.ConverterUtils.DataSource;
public class RiversClassification
{
public static void main(String args[]) throws Exception{
DataSource source = new DataSource("C:\\Users\\Bindra\\Desktop\\data set\\-
claasification\\train2008.arff");
Instances trainDataset = source.getDataSet();
trainDataset.setClassIndex(trainDataset.numAttributes()-1);
int numClasses = trainDataset.numClasses();
for(int i = 0; i < numClasses; i++){
String classValue = trainDataset.classAttribute().value(i);
System.out.println("Class Value "+i+" is "+ classValue);
}

//Classifier used is Naïve Bayes here
NaiveBayes nb = new NaiveBayes();
nb.buildClassifier(trainDataset);
DataSource source1 = new DataSource("C:\\Users\\Bindra\\Desktop\\data set\\-
claasification\\test2011.arff");
Instances testDataset = source1.getDataSet();
testDataset.setClassIndex(testDataset.numAttributes()-1);
System.out.println("=====");
System.out.println("Actual,NB Predicted Class");
for (int i = 0; i < testDataset.numInstances(); i++) {
double actualClass = testDataset.instance(i).classValue();
String actual = testDataset.classAttribute().value((int)actualClass);
Instance newInst = testDataset.instance(i);
double predNB = nb.classifyInstance(newInst);
String predString = testDataset.classAttribute().value((int) predNB);
System.out.println(actual+", "+predString);
}
}
}

```

4 Performance Comparison

Below is the performance of various classification techniques on the used data set (Table 1).

Table 1 Percentage error in classification techniques applied for analysis of water quality of rivers

Classification technique	Incorrectly classified elements (%)
Naïve Bayes	46.667
J48	60
SMO	66.667
REPTree	73.333

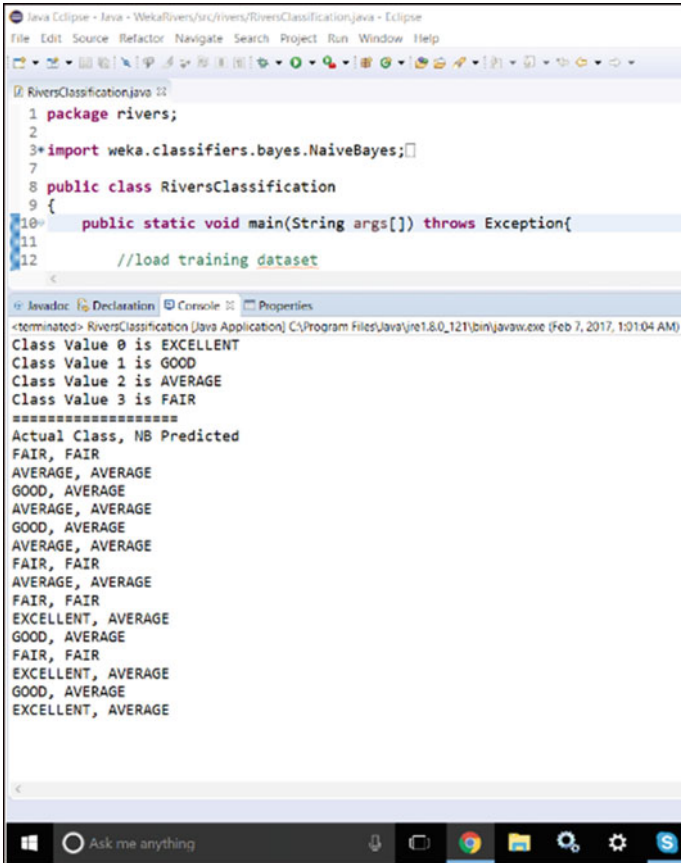
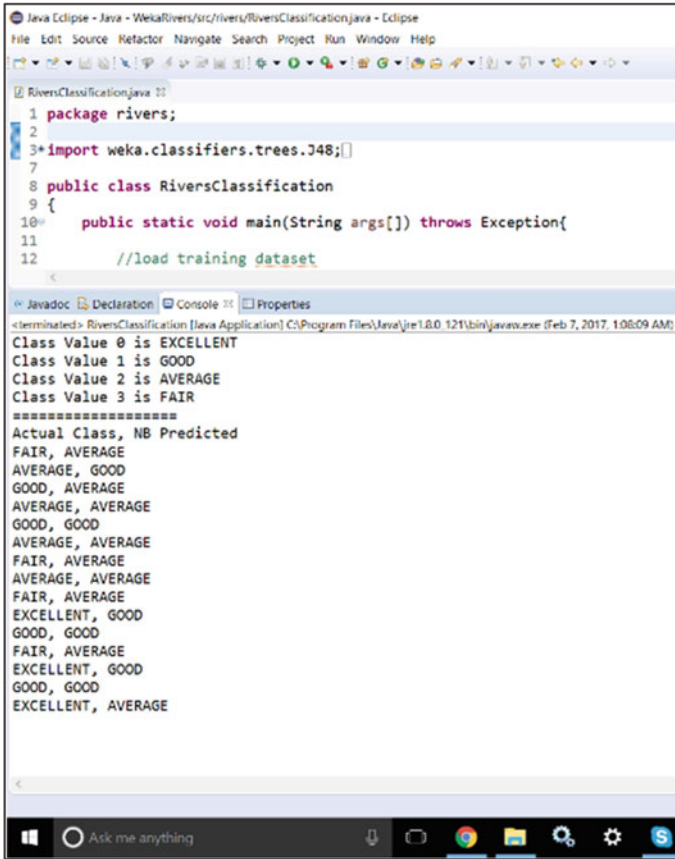


Fig. 1 Output of the code for Naïve Bayes classifier

The table consists of two columns, techniques and error percentage. Naïve Bayes show the best result and maximum error is found in REPTree technique.

Below are the screenshots of the outputs when the code was run on Eclipse IDE using different classifiers (Figs. 1, 2, 3, and 4).



```
Java Eclipse - Java - WekaRivers/src/rivers/RiversClassification.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help

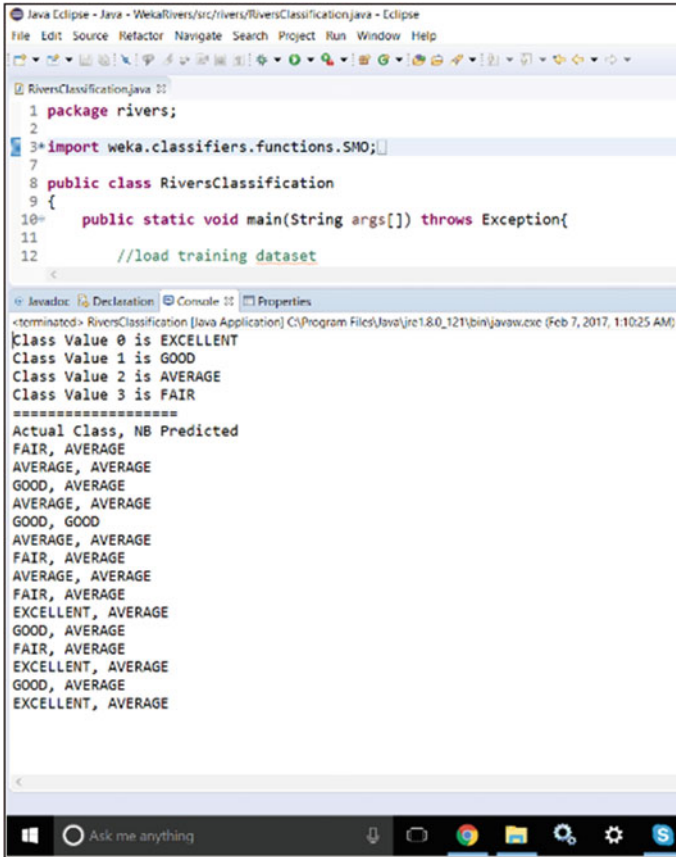
RiversClassification.java
1 package rivers;
2
3 import weka.classifiers.trees.J48;
4
5 public class RiversClassification
6 {
7     public static void main(String args[]) throws Exception{
8         //load training dataset
9     }
10
11
12

Javadoc Declaration Console Properties
<terminated> RiversClassification [Java Application] C:\Program Files\Java\jre1.8.0_121\bin\javaw.exe (Feb 7, 2017, 1:08:09 AM)
Class Value 0 is EXCELLENT
Class Value 1 is GOOD
Class Value 2 is AVERAGE
Class Value 3 is FAIR
=====
Actual Class, NB Predicted
FAIR, AVERAGE
AVERAGE, GOOD
GOOD, AVERAGE
AVERAGE, AVERAGE
GOOD, GOOD
AVERAGE, AVERAGE
AVERAGE, AVERAGE
FAIR, AVERAGE
AVERAGE, AVERAGE
FAIR, AVERAGE
EXCELLENT, GOOD
GOOD, GOOD
FAIR, AVERAGE
EXCELLENT, GOOD
GOOD, GOOD
EXCELLENT, AVERAGE
```

Fig. 2 Output of the code for J48 classifier

5 Result and Discussion

In this work, Naïve Bayes has proved to be the best technique with minimum error. The error percentage in the classification techniques we have applied in our analysis is high because the data set which is used as input is biased since in the dataset



```
Java Eclipse - Java - WekaRivers/src/rivers/RiversClassification.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help

RiversClassification.java ::
1 package rivers;
2
3 *import weka.classifiers.functions.SMO;
7
8 public class RiversClassification
9 {
10= public static void main(String args[]) throws Exception{
11
12 //load training dataset

-----
<terminated> RiversClassification [Java Application] C:\Program Files\Java\jre1.8.0_121\bin\javaw.exe (Feb 7, 2017, 1:10:25 AM)
Class Value 0 is EXCELLENT
Class Value 1 is GOOD
Class Value 2 is AVERAGE
Class Value 3 is FAIR
=====
Actual Class, NB Predicted
FAIR, AVERAGE
AVERAGE, AVERAGE
GOOD, AVERAGE
AVERAGE, AVERAGE
GOOD, GOOD
AVERAGE, AVERAGE
FAIR, AVERAGE
AVERAGE, AVERAGE
FAIR, AVERAGE
EXCELLENT, AVERAGE
GOOD, AVERAGE
FAIR, AVERAGE
EXCELLENT, AVERAGE
GOOD, AVERAGE
EXCELLENT, AVERAGE
```

Fig. 3 Output of the code for SMO classifier

number of instances in average and good water quality groups are greater than the number of instances in fair and excellent groups. Another reason for high error percentage is that the number of instances in excellent, good, fair and average water quality groups are not same.

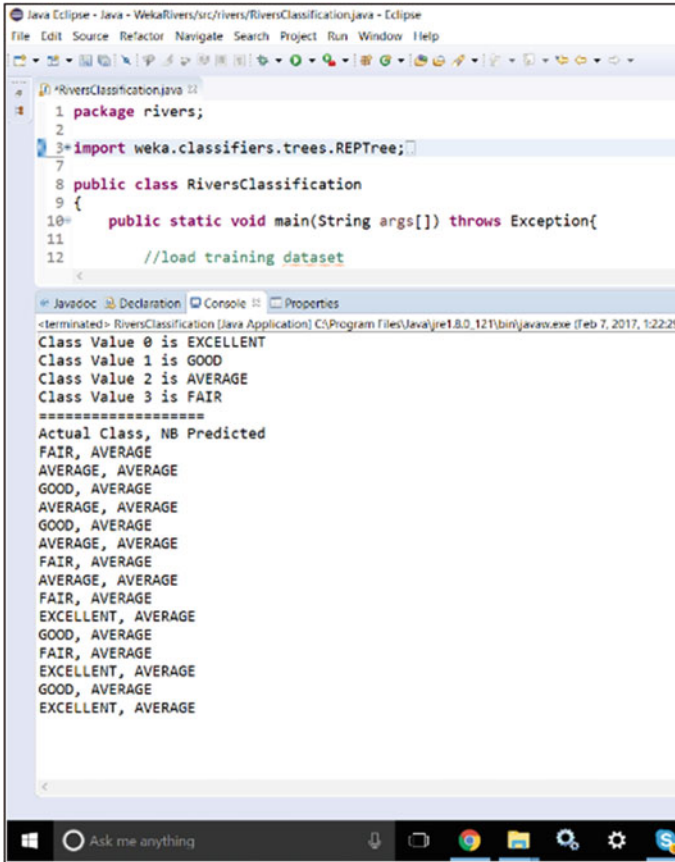


Fig. 4 Output of the code for REPTree classifier

6 Conclusion

Out of all the classification techniques, we have applied on our dataset, Naive Bayes has shown the results with least error.

In future, for effective and accurate analysis, some modifications need to be applied in these predefined classification techniques or new classification techniques need to be devised in order to form correct classes of such biased datasets.

Acknowledgements I profoundly thank Bharati Vidyapeeth's College of Engineering for constant support and encouragement.

References

1. Masethe, H. D., & Masethe, M. A. (2014). Prediction of heart disease using classification algorithms. In *Proceedings of the World Congress on Engineering and Computer Science 2014* (Vol. II), October 22–24, 2014, San Francisco, USA.
2. Saini, P., & Jain, A. K. (2013). Prediction using classification technique for the students' enrollment process in higher educational institutions. *International Journal of Computer Applications* (0975–8887), 84(14).
3. Padmapriya, A. Dr. (2012). Prediction of higher education admissibility using classification algorithms. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(11).
4. Sudhakar, K., & Manimekalai, M. Dr. (2014). Study of heart disease prediction using data mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(1).
5. Srivastava, G., & Kumar, P. (2013). Water quality index with missing parameters. *IJRET: International Journal of Research in Engineering and Technology*, 02(04), 609–614.
6. National River Conservation Directorate (NRC) <http://envfor.nic.in/division/national-river-conservation-directorate-nrcd>. Date accessed on 30/9/2016.
7. Data Set <https://data.gov.in/catalog/status-water-quality-india-2008-and-2011>.
8. NDASP <http://www.dst.gov.in/national-data-sharing-and-accessibility-policy-0>.
9. Ministry of Environment and Forests <https://data.gov.in/ministrydepartment/ministry-environment-and-forests>.
10. Sujatha, M., Prabhakar, S., & Lavanya Devi, G. Dr. (2013). A survey of classification techniques in data mining. *International Journal of Innovations in Engineering and Technology (IJJET)*, 2(4). ISSN 2319-1058.
11. Bhargavi, P., & Jyothi, S. Dr. (2009). Applying Naive Bayes data mining technique for classification of agricultural land soils. *IJCSNS International Journal of Computer Science and Network Security*, 9(8).
12. Patil, T. R., & Sherekar, S. S. Mrs. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, 6(2). ISSN 0974-1011.
13. Platt, J. C. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines* (Technical Report MSR-TR-98-14), April 21, 1998.
14. Kalmegh, S. (2015). Analysis of WEKA data mining algorithm REPTree, simple cart and RandomTree for classification of Indian News. *IJISSET—International Journal of Innovative Science, Engineering & Technology*, 2(2). ISSN 2348–7968.
15. Eclipse IDE <http://www.eclipse.org/users/>. Date accessed on 11/2/2017.
16. Weka website (Latest version 3.6) <http://www.cs.waikato.ac.nz/ml/weka/>. Date accessed on 30/9/2016.