International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Data Mining Case Study for Water Quality Prediction using R Tool

S. Pavithra Devi[1], S. Jothi[2] & A. Devi[1]
[1]M. Sc. (CS&IT), Jayaraj Annapackiam College for Women, Periyakulam, Tamil Nadu, India
[2]Assistant Professor of Computer Science, Jayaraj Annapackiam College for Women, Periyakulam, Tamil Nadu, India

ABSTRACT

Water pollution has a duel effect on nature. It has negative effects on the living and also on the environment. The effects of pollution on human beings and aquatic communities are many and varied. Water pollution causes approximately 14,000 deaths per day, mostly due to contamination of drinking water by untreated sewage in developing countries. Human activities including industrialization and agricultural practices contributed immensely in no small measure to the degradation and pollution of the environment which adversely has an effect on the water bodies (rivers and ocean) that is a necessity for life. Pollution is the introduction of a contamination into the environment. It is created by industrial and commercial waster, agricultural practices, everyday human activities and most notably, models of transportation. This paper tries to discuss basically what water pollution is and equally to address the source, effect control and water pollution management as a whole. Some recommendations such as introduction of environmental education were mentioned.

Keywords : Pollution, Contamination, Water recycling, Dengue, Maleria and Chicenguniya

I. INTRODUCTION

The importance of water for sustenance of life cannot be overemphasized. Whether it is in use of running water in our homes, rearing cattle and growing crops in our farms, or the increased uses in industry, remain immeasurable. It is important therefore, to not that depletion of this commodity either through contamination, or careless use results in serious consequences.

Water is the most vital element among the natural resources, and is critical for the survival of all living organisms including human, food production, and economic development. Today there are many cities worldwide facing an acute shortage of water and nearly 40 percent of the world's food supply is grown under irrigation and a wide variety of industrial processes depends on water. The environment, economic growth, and developments are all highly influenced by water-its regional and seasonal availability, and the quality of surface and groundwater. The quality of water is affected by human activities and is declining due to the rise of urbanization, population growth, industrial production, climate change and other factors. The resulting water pollution is a serious threat to the well-being of both the Earth and its population.

Water is considered polluted if some substances or condition is present to such a degree that the water cannot be used for a specific purpose. Olaniran (1995) defined water pollution to be the presence of excessive amounts of a hazard (pollutants) in water in such a way that it is no long suitable for drinking, bathing, cooking or other uses. Pollution is the introduction of a contamination into the environment. It is created by industrial and

commercial waster, agricultural practices, everyday human activities and most notably, models of transportation. No matter where you go and what you do, there are remnants earths environmental and its inhabitants in many ways. Water pollutants could be:

a. Biological (pathogens, such as viruses, bacteria, protozoa, algae and helminths),
b. Chemical (organic chemicals, like biocides, polychlorinated biphenyls or PCBs; inorganic chemicals, like phosphates, nitrates, fluoride, etc., also heavy metals like As, Pb, Cd, Hg, etc.,

Sediment pollution - according to the type of particular available in water it could be which is caused by soil particles that enter the water as a result of runoff from agricultural lands, forests, over grazed rangelands, strip mines and construction sites or Sewage pollution caused by Liquid wastes from domestic activities such as kitchen, toilet and other household wastewater.

II. RESEARCH METHODOLOGY

2.1 Methodology

The present study is an exploratory research conducted among the people in Dindigul. In order to pursue the aims and objectives outlined in the introduction, a content analysis of information gained from a multimedia research process was conducted to establish the underlying trends in location to find common diseases.

The first stage involved gathering of secondary information from people. The second stage involved identifying the age group among them and structuring a comparative analysis of the five identified parameters under each category.

A summary of interpretations was also given. In the third stage, analysis was carried out by making specific assumptions in a hypothetical situation. In the last and the fourth stage, on the basis of the results and interpretations, specific postulates were framed, and on each postulate hypotheses were framed that can be tested through quantitative research in the future. The above-mentioned stages have been described as objectives in the preceding paragraph.

2.2 Algorithm Used

Data mining is the core process of knowledge discovery in database. It is the process of extraction of useful patterns from the large database. To analyze the large amount of collected information, the area of Knowledge Discovery in Database (KDD) provides techniques which extract interesting patterns in a reasonable amount of time. Data mining is the application of efficient algorithms to detect the desired patterns contained within the given data. Data mining is the extraction of hidden descriptive or predictive information from large databases.

Association Rule Mining

Association rules mining are one of the major techniques of data mining. The purpose of association analysis is to figure out the hidden association and some useful rules of data base, and uses these rules to speculate and judge the unknown matter from the already known information. Association rule mining has many important applications in our life.

Association Rule

An association rule is one of the forms x=>y. and each rule has two basic needs: support and confidence. Things that occur often together can be associated to each other. Conclusions based on the frequent item sets make association rules.

2.2.1 APRIORI ALGORITHM

APRIORI algorithm is a fundamental algorithm mining association rule. It contains two processes:

- Detect all frequent item sets by scanning db.
- Form strong association rules in the frequent item sets.

Process one needs to scan DB several times, which consumes a lot of time and space. As a result, what needs to be improved is the mining competency of frequent group of things in DB. Apriori algorithm is a significant algorithm for mining frequent itemsets for Boolean association rules. Apriori algorithm is formed by Agrawal and Srikantin 1994. It is the most fundamental and important algorithm for mining frequent itemsets. Apriori is used to detect all frequent itemsets in a provided database db.Thekeynote of Apriori algorithm is to form multiple passes over the database. It employs a repetitive approach called as a breadth-first search (level-wise search).

2.2.2 Key Concepts

- **Frequent Item Sets:** The item sets which has minimum help (denoted by li for i^{th}-item sets), Apriori property: any subgroup of frequent things must be frequent.
- **Join Operation:** to detect lk, a group of candidate k- group of things is developed by adding lk-1with itself.

How Apriori Works?

- **Find All Frequent Item sets.**
- **Get Frequent Things**: Things whose occurrence in database is more than or equal to the minimum help threshold.
- **Frequent Item Sets**: Develop candidates from frequent things. Prune the results to detect the frequent item sets. Develop strong association rules from frequent item sets. Rules which satisfy the minimum support and minimum confidence threshold.
- **Association Rule**: Association rule of data mining involves picking out the unknown inter-dependence of the data and finding out the rules between those items [3]. Agrawal introduced association rules for point of sale (POS) systems in supermarkets. A rule is defined as an implication of the form A=>B, where A∩ B≠Ø. The left-hand side of the rule is called as antecedent. The right-hand side of the rule is called as consequent.
- **Support:** I = { i_1,i_2,i_3, … , i_m} is a collection of items. T be a collection of transactions associated with the items. Every transaction has an identifier TID [6]. Association rule A=>B is such that A Є I, B Є I. A is called as Premise and B is called as Conclusion. The support , S, is defined as the proportion of transactions in the data set which contains the item set.

Support(X=>Y) = Support (XUY) = P(XUY)

- **Confidence:** The confidence is defined as a conditional probability Confidence (X=>Y) = Support (XUY) / Support(X) = P(Y/X). Lift: is the ratio of the probability that L and R occur together to the multiple of the two individual probabilities for L and R, i.e. lift = Pr(L,R) / Pr(L).Pr(R).
- **Conviction:** is similar to lift, but it measures the effect of the right-hand-side not being true. It also inverts the ratio. So, a conviction is measured as:

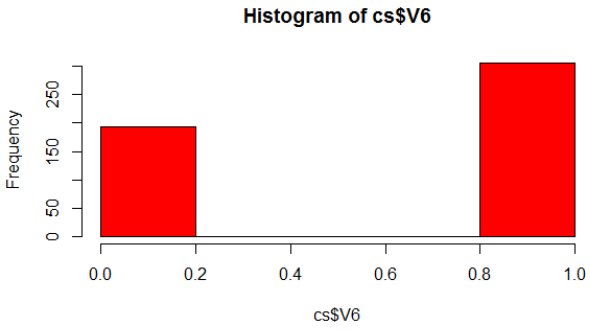
Conviction = Pr(L).Pr(not R) / Pr(L,R)

III. FINDINGS AND INTERPRETATION

3.1 Findings and Interpretations

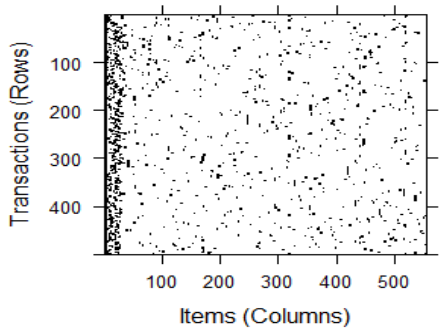
Histogram:

> hist(cs$V4)



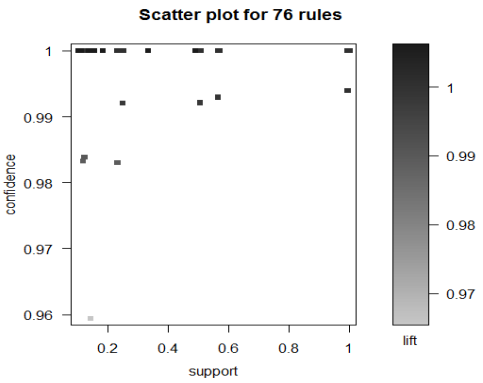
Dotchart:

> plot(rules, method = "grouped")



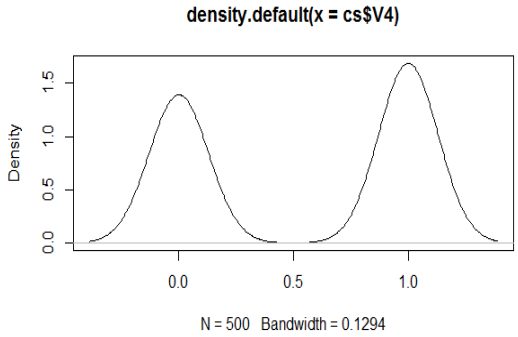
Scatter Plot:

> plot(rules)



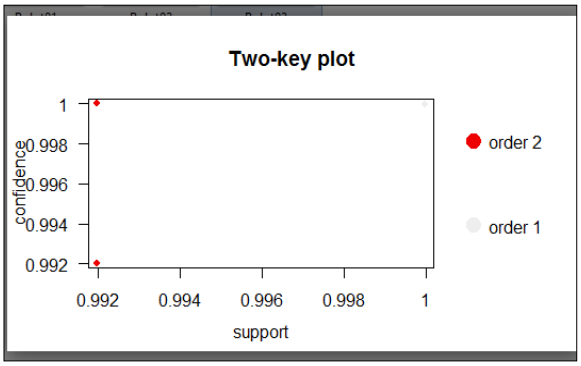
Density:

> d<-density(cs$V4)



Two Key Plot:

> plot(rules,shading="order",
control=list(main="Two-keyplot"));

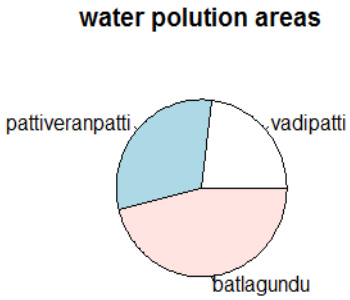


Pie Chart:

slices<-c(6,8,12)

> lbls<-c("vadipatti","pattiveranpatti","batlagundu")

> pie(slices,labels = lbls,main="water polution areas")



IV. CONCLUSION

In this paper an analysis is presented for water quality prediction using various data mining techniques at different locations. Clean water is a limited resource and hence its protection and quality running is highly essential for sustainable development. Fresh water crisis in Tamil Nadu, especially in metropolitan areas is growing day by day and is a future warning to water sustainability. In this paper, research has conducted experiments for assessing the water quality of the areas in Batlagundu, Vadipatti and Pattiveeranpatti. The study has bridge the gap in pollution data analysis and shows the potential of the apriori algorithm. The histogram diagram express that Batlagundu area people are using polluted water at the frequency of 125 to 155. Totaly 3 namely Vadipatti, Pativeeranpatti and area people are considered for this research.

The density of the dataset varies from 0.1 to 1.4 and it's having the bandwidth of 0.1294. This analysis revealed more intresting patterns. The Pie Chart give the conclusion that Baltagundu people using more polutted water as compared to other areas like Vadipatti and Pattiveeranpatti people. So Batlagundu people having the possibility of affecting Dengue, Maleria and Chicenguniya, than other area people.

V. SUGGESTIONS

There are several promising directions to extend the work presented is this case study.

- ✓ This research can be extended to particlar city and can find different reasons to and solutions to clean the water resources.
- ✓ Assess the prevalent health conditions of the people living around water and prepare a health profile.
- ✓ It is recommended that there should be proper waste disposal system and waste should be treated before entering in to river.
- ✓ Educational and awareness programs should be organized to control the pollution.
- ✓ Water recycling and wastewater management can be implemented in urban cities for water sustainability

VI. REFERENCES

[1]. P. H. McGauhey, Engineering Management of Water Quality, 1968.

[2]. H. Peavy, D. Rowe, and G. Tchobanoglous, Environmental Engineering, 1986.

[3]. M. Abedin, Health and Population Sector: An Overview and Vision, in Logical Framework (Log-Frame) Workshop for the Fifth Health and Population Programme (HAPP-5), pp. 23-25, 1997.

[4]. S. Ahmed, K. Tapley, A. Clemett, and M. Chadwick, Health and Safety in the Textile Dyeing Industry, 2005.

[5]. K. Akhter, Studies on Water Quality in the Peripheral River System Around Dhaka City, 2007.

[6]. M. Ahmed, Sector Review Industry, Environmental Management Training Project Instructor, 1993.

[7]. M. Ahmed and M. Rahman, Water Supply and Sanitation, 2000.

[8]. D. Bhattacharya, B. Kabir, and K. Ali, Industrial Growth and Pollution in Bangladesh : A sectoral Analysis, in Symposium on Environment and Sustainable Development with Special Reference to Bangladesh, North South University, Dhaka, 1995.

[9]. BBS, Report of Health and Demographic Survey, 2000.

[10]. G. Browder, Final Report of Water Quality Management Task (ADBTA1104-BAN), National Environmental monitoring and Pollution Control Project, The Asian Development Bank, 1992.