International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-5, March 2019

Quality Assessment of Ground Water on Small Dataset

Aiswarya Vijayakumar, A S Mahesh

Abstract*: Quality assessment of water has a lot of attractions during recent years. Diverse kinds of classification and monitoring techniques were used in this field of study. The present examination investigates the quality of ground water in Kudankulam which is situated Tirunelveli district of Tamil Nadu. A total of 19 samples was accumulated in this region typically from the coastal area during 2011-2012.The evaluation was done on the basis chemical parameters of each samples. This paper explores various classifier models such as KNN, NB and SVM to achieve prediction of groundwater quality. The classification is done based on the Water Quality Index (WQI) of each sample. A near investigation of characterization systems was done dependent on the confusion matrix, accuracy, f1 score, precision and recall. The outcomes propose that SVM is a better method having high accuracy rate than other models.

***Index Terms*: Classification Algorithms, Water Quality Index, Support Vector Machine, Naïve Bayes; K-Nearest Neighbors.**

I. INTRODUCTION

Nowadays, use of groundwater extended at an aggravating amount over the world, since it can be obtained from various sources. Due to this the abuse of ground water also increases at a tremendous amount, which badly affects the surface water resources. Hence, it is important to process the ground water to determine the quality for current and future consumption. There exists a huge amount of geochemical works are done in this field to discover the reasonableness of groundwater. WQI is a helpful tool for quality analysis of water [14]. In this field, classification is helpful method to determine the quality of water ie, either good or bad. Classification of information into various classes is a strategy to arrange extensive information for efficient computation. Supervised methods are progressively utilized in a few fields, for example, medical sciences, pharmaceutical science, and social and monetary sciences [2]. Comparison of different classifications was carried out in the basis of Electrical conductivity levels from the samples collected Madhya Pradesh, India. The implementation process was done with the help of classification learner application in MATLAB [8]. An evaluation was done on 41 samples that collected from Khuzestan Province, Iran with 17 parameters. The interpretation is done the basis of pollution level readings, SVM and K-NN classifier models used for the analysis of WQ [10].This paper focuses on a comparative study of different classification techniques such as KNN, NB and SVM for quality assessment of ground water. The WQI is the key factor for classification. Based on variations of each chemical/mineral components the WQI also differs. That is WQI value for each sample will be unique. The model is evaluated on the basis of precision, confusion matrix, recall and accuracy.

The methodology adopted in this paper is as follows:

- Data Procuring
- Data Preprocessing
- Model Implementation
- Model Evaluation

Manuscript published on 30 March 2019.
***Correspondence Author(s)
Aiswarya Vijayakumar, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India.
A S Mahesh, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India

II. BACKGROUND STUDY

Classification techniques are widely adopted in the field of Environmental Sciences. There exist a lot of investigations that tells us these techniques are very much useful for water quality research. The procedure that adopted in classification is to categorize data for efficient computation. Description of algorithms used in study is as follows:

A. K-Nearest Neighbors

KNN is fast and simple algorithm that used to classify data based on nearest neighbors by calculating distance functions [3]. The commonly used distance function is Euclidean. The other functions are Minkowski , Manhattan or Hamming distance.

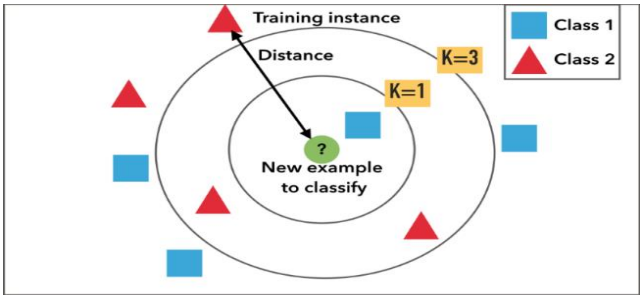
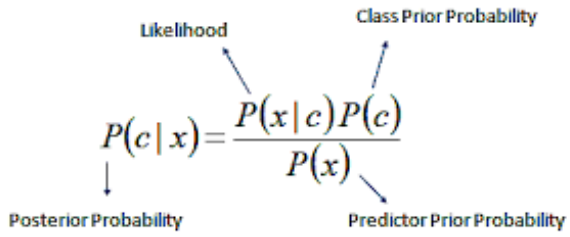


Figure 1: Basic illustration of KNN

B. Naïve Bayes

NB is commonly used, effective classification model. It mainly works on the basis of probability, by using Bayes Theorem. NB is popularly used in fields of recommendation systems, text classification etc. it is easy to implement on a small dataset [9]. It can be represented using Bayesian network. The equation used in NB is showed in figure 2

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

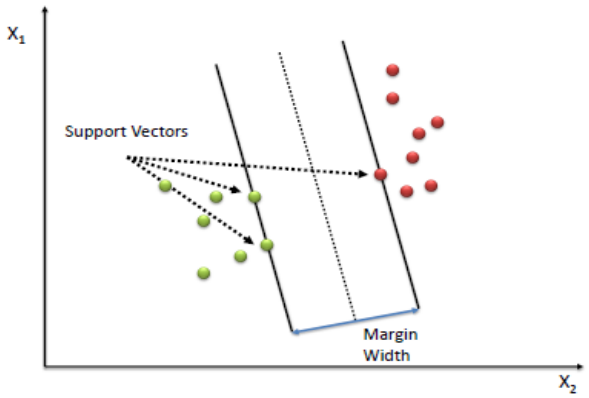


$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 2: Basic illustration of equation used in NB

C. Support Vector Machine

SVM is the most common and successful machine learning algorithm can be used for both regression and classification .It is used to maximize the accuracy rate using different kernel tricks with minimizing over fitting of data. A proper SVM gives us a hyper plane that correctly divides the data into two classes without overlapping. Foremost thing in SVM model is to choose hyper plane immense possible margin and training set with data points (support vectors). To minimizing the error the data should be classified correctly with right hyper plane [4].



. Figure 3: Illustration of basic SVM

Commonly used kernel functions in SVM shows as follows:

Polynomial K $(\mathbf{X_m}, \mathbf{X_n})$ $= (\gamma \mathbf{Xm.X_n} + C)^d$ *(1)*

RBF K$(\mathbf{X_m}, \mathbf{X_n})$ $= \exp(-\gamma|\mathbf{X_m} - \mathbf{X_n}|^2)$ *(2)*

Linear K $(\mathbf{X_m}, \mathbf{X_n})$ $= \mathbf{X_m.X_n}$ *(3)*

Sigmoid K $(\mathbf{X_m}, \mathbf{X_n})$ $= \tanh(\gamma|\mathbf{X_m} .\mathbf{X_n} + C)$ *(4)*

Where K$(\mathbf{X_m}, \mathbf{X_n}) = \phi(\mathbf{X_m}) .\phi(\mathbf{X_n})$

III. METHODOLGY

A. Data Procuring

The data used for this study have been collected from Kudamkulam, Tamil Nadu. Totally 19 samples were collected typically from coastal area during the year 2011-2012.The data consists with parameter as mineral components.

B. Data Preprocessing

In this study the preprocessing step that involves removal of noisy data and the parameters gets selected on the basis of permissible and desirable limit that suggested by Bureau of Indian Standards (BIS).The samples processed using the chemical parameters and quality rating(QR) to calculate WQI. In this study WQI for each sample calculated separately, based on the value of WQI the target class assigned for the classification. The Horton [5] method is used to evaluate WQI for each sample. Since it is a binary classification the status of water quality (wq) will be either fit or unfit. The succeeding equation confess calculate WQI [14].

$WQI = \Sigma Q_i W_i / \Sigma W_i$ *(5)*

Q_i = QR of i^{th} WQ parameter.

W_i = UW of i^{th} WQ parameter, UW = Unit Weight

Table 1: WQI calculation for sample 1

Sl. No.	Parameters	Observed Values	Standard Values (S_n)	Unit Weight (W_n)	Quality Rating (q_n)	W_nq_n
1	pH Value	7.79	8.5	0.0978	158	15.46
2	Total Hardness	183.87	200	0.0042	91.94	0.38
3	Calcium	28.15	75	0.0111	37.53	0.42
4	Magnesium	73.89	30	0.0277	246.30	6.83
5	Chlorides	199	250	0.0033	79.60	0.26
6	Total Dissolved Solids	1510	500	0.0017	302	0.50
7	Sulphate	6.7	200	0.0042	3.35	0.01
				∑Wn = 1.000	∑qn = 937.16	∑qn Wn = 32.34
Water Quality Index = ∑qn Wn / ∑ Wn = 32.34						

Accordingly WQI for each sample can calculated as mentioned in table 1

C. Model Implementation

The implementation of 3 models was done in PYTHON 3.6. For this purpose dataset was divided as 75% for training and 25% for testing. The data is classified on basis of WQI value of each sample; since the rate will be differ from one to other samples. The NB, KNN and SVM models are implemented on ground water datasets. The result of this interpretation will be either fit or unfit based WQI rate.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The result divides into two categories that is, 0-25 good water quality and 26-100 bad water quality. Figure 4 shows the work flow of this study.

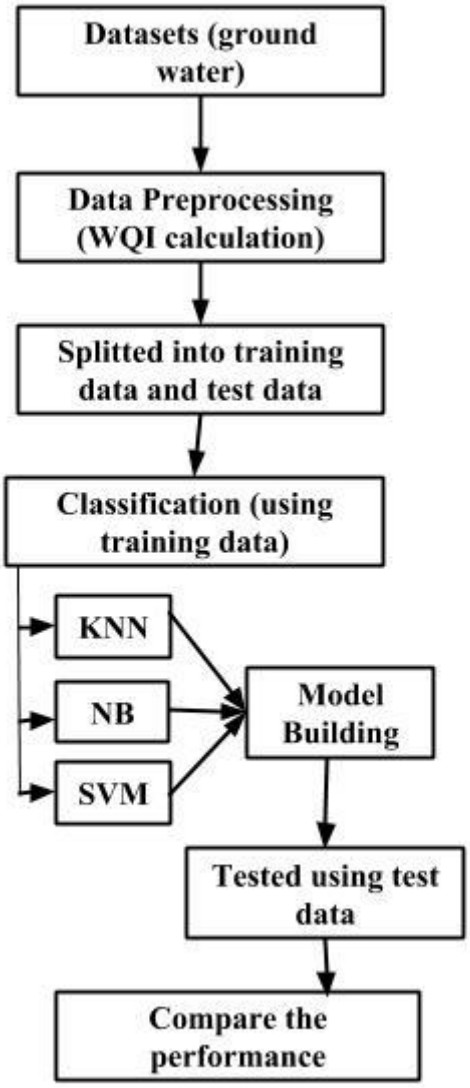


Figure 4: Work flow of this study

D. Model Evaluation

In Machine learning, the performances of models analyzed with the help of Confusion Matrix (CM).It is a table that consist of 4 distinct mixes of genuine and anticipated qualities of data that are testing [13]. It consist of positives of true (T^+), negatives of true (T^-) , positive of false (F^+) and negatives of negative (F^-) respectively. The performance of each models assessed on the basis of Precision (P), Accuracy (A), Recall(R), and F1 Score (F1). Equations (6 – 9) explain how to calculate these metrics.

$$A = (T^+ + T^-)/ (T^+ + T^- + F^+ + F^-) \quad (6)$$

$$P = T^+ / (T^+ + F^+) \quad (7)$$

$$R = T^+ / (T^+ + F^-) \quad (8)$$

$$F1 = 2 *(R*P) / (R+P) \quad (9)$$

IV. RESULTS AND DISCUSSIONS

The presented work implemented in PYTHON 3.6. The result evaluated using 70:30 ratio for training and testing respectively. The result gained after testing the model using test data with different classification models showed in Table 2.

Table 2: Performance of models with overall accuracy

Models	Accuracy
KNN	66
NB	83
SVM	**98**

The table 2 clearly shows that the SVM model gives better result than other models. In NB, gives the same result for both Gaussian and Multinomial methods. KNN gives minimum accuracy in this study.

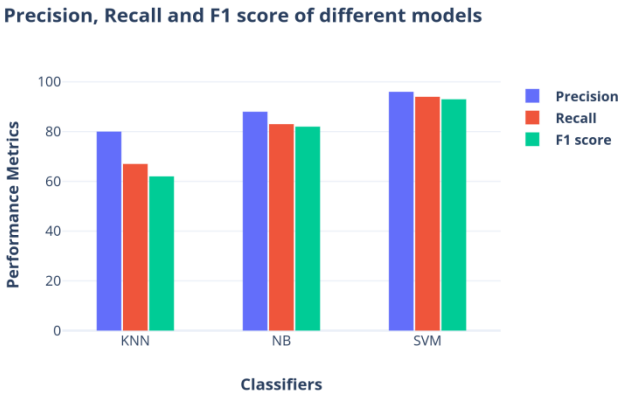


Figure 5: Graphical representation of performance metrics of different models.

The Fig 5 shows recall, f1 score and precision of different models in small dataset. It explains that SVM gives highest performance rate than other models.

V. CONCLUSION

Environmental science is a trending field that widely used for classification and prediction in these days. In that, water quality monitoring is the key area because it is predominant for nature. In this interpretation WQI is the classification criteria used to assess the wq of samples. The samples amassed from kudankulam of Tamil Nadu.

This study implemented in PYTHON 3.6 and result evaluated using CM, F1, P, R and Accuracy separately for each models. The study clearly depicts that SVM is the best model for WQ assessment on small dataset. The performance metrics for SVM is higher than other models (KNN, NB).

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

FUTURE WORK

This study utilized KNN, NB and SVM technique. For future Investigations ANN can be utilized with multi variable conditions and other strategies like deep learning, LDA can also use.

ACKNOWLEDGMENT

The authors gratefully acknowledge Dr Hudson Oliver for providing genuine data to carry out this study.

REFERENCES

1. Central Ground Water Board, Ministry of Water Resources,Government of India- BIS Standard
2. Dollar, E. S. J., James, C. S., Rogers, K. H., &Thoms, M. C. (2007). A framework for interdisciplinary understanding of rivers as ecosystems. *Geomorphology*, *89*(1-2), 147-162.
3. Dubey, H. (2013). Efficient and accurate kNN based classification and regression. *A Master Thesis Presented to the Center for Data Engineering, International Institute of Information Technology, Hyderabad-500*, *32*.
4. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., &Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.
5. Horton, R. K. (1965). An index number system for rating water quality. *Journal of Water Pollution Control Federation*, *37*(3), 300-
6. Khalil, A., Almasri, M. N., McKee, M., &Kaluarachchi, J. J. (2005). Applicability of statistical learning algorithms in groundwater quality modeling. *Water Resources Research*, *41*(5).
7. Moore, A. W. (2001). Support vector machines. Tutorial. School of Computer Science of the Carnegie Mellon University. Available at http://www. cs. cmu. edu/~ awm/tutorials
8. Prakash, R., Tharun, V. P., & Devi, S. R. (2018, April). A Comparative Study of Various Classification Techniques to Determine Water Quality. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1501-1506).
9. Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).
10. Sakizadeh, M., &Mirzaei, R. (2016). A comparative study of performance of K-nearest neighbors and support vector machines for classification of groundwater. *Journal of Mining and Environment*, *7*(2), 149-164.
11. Sakizadeh, M. (2015). Assessment the performance of classification methods in water quality studies, A case study in Karaj River. *Environmental monitoring and assessment*, *187*(9), 573.
12. Srinivas, Y., Hudson, O. D., Stanley, R. A., &Chandrasekar, N. (2014). Quality assessment and hydrogeochemical characteristics of groundwater in Agastheeswaram taluk, Kanyakumari district, Tamil Nadu, India. *Chinese Journal of Geochemistry*, *33*(3), 221-235.
13. Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, *9*(1), 40-50.
14. Yogendra, K., &Puttaiah, E. T. (2008). Determination of water quality index and suitability of an urban waterbody in Shimoga Town, Karnataka. In *Proceedings of Taal2007: The 12th world lake conference* (Vol. 342, p. 346).
15. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.
16. Aiswarya Vijayakumar, A S Mahesh (2019). Quality Assessment of Groundwater in Pre and Post Monsoon using various classification Techniques (Unpublished).

AUTHORS PROFILE

Aiswarya Vijayakumar is a MPhil Scholar at Department of Computer Science and IT, in Amrita School of Arts and Sciences Kochi, India. She completed her BCA from Bharathiar University, Tamil Nadu and Post-graduation (MCA from Amrita Vishwa Vidyapeetham, Tamil Nadu. Her interested areas includes Data Mining, Machine learning, Hadoop and Cloud Computing.

A S Mahesh is a Assistant Professor in Department of Computer Science and IT, in Amrita School of Arts and Sciences Kochi, India. His qualifications include M. Sc. (CS), MBA (Systems and Marketing), and M.Phil. (CS). He has more than 19 years of teaching including 6 years in research. His areas of interest includes Cloud Computing, Networking and Programming.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)