WATER QUALITY PREDICTION

REVIEW OF LITERATURE

Machine Intelligence

BACHELOR OF TECHNOLOGY

Department of Computer Science & Engineering

V Semester Section: B

SUBMITTED BY

Batch No: 19

NAME	SRN
Ayush Singh	PES2UG20CS080
Ayan Aggarwal	PES2UG20CS079

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)

100 Feet Ring Road, BSK III Stage, Bengaluru-560085

PAPER 1: Water quality prediction method based on LSTM neural network.
Authors:Yuanyuan Wang, Jian Zhou, Kejia Chen, Yunyun Wang, Linfeng Liu

A new water quality prediction method based on long and short term memory neural networks (LSTM NN) for water quality prediction is proposed in this paper. Firstly, a prediction model based on LSTM NN is established. Secondly, as the training data, the data set of water quality indicators in Taihu Lake which measured monthly from 2000 to 2006 years is used for the training model. Thirdly, to improve the predictive accuracy of the model, a series of simulations and parameters selection are carried out. Finally, the proposed method is compared with two methods: one is based on a back propagation neural network, the other is based on an online sequential extreme learning machine. The results show that the method is more accurate and more generalized.

The long and short memory neural network (LSTM NN) has 'memory' because of its own unique network structure. It has been applied in the field of time prediction successfully such as stock prediction [5] and traffic flow prediction [6].

Water quality indicators are the time series data and greatly affected by seasons with obvious seasonal diversity. Water quality prediction belongs to the time series prediction. Because the traditional neural network isn't suitable for dealing with the time series data, this paper proposes a water quality prediction method based on LSTM NN.

In order to make an accurate prediction of water quality indicators, we establish a model based on LSTM NN for water quality prediction as shown in Fig. 3. Compared with the traditional neural network, the nodes in the hidden layer are fully connected and adopt the structure of the memory block.

In order to improve the predictive accuracy of the model, the model is trained by historical data of water quality indicators. In training, the model selects the best performance with different time step (*d)*, the number of iterations (*epoch)* and the number of neurons in hidden layer (*Hidd-num)*.

In order to improve the predictive accuracy of the model, the model makes a parameter selection among time step (*d)*, the number of iterations (*epoch)* and the number of neurons in the hidden layer (*Hiddnum)*. In this paper, several simulations are carried out

To compare results obtained from LSTM NN, BP NN and OS-ELM, the same test set is used. When the fluctuation of data is small, the predictive accuracy of BP NN is higher than OS-ELM. When the fluctuation of data is large, the predictive accuracy of OS-ELM is higher than BP NN. The predictive accuracy of LSTM NN is always higher than the first two in both cases. LSTM NN has been relatively stable which always be the smallest with each time step. All experimental results demonstrate that LSTM NN is generalized and effective for the prediction of water quality indicators.

Besides, LSTM NN is more generalization. Considering the disadvantage of long training cycle, To improve the predictive accuracy of the model, several simulations and parameter selection are carried out. a more effective memory block will be designed in the future work.

PAPER 2: Indicator analysis of lake water quality based on fuzzy neural network and spectrophotometry.

Authors: Yinshan Yu, Yan Qu, Hongyun Zhang, Lingjie Jiang, Mingzhen Shao, Dandan Wei, Dongyang Zhang

This paper proposes a method of using fuzzy neural network modelling to predict water quality indicators to handle the difficult problems of detecting lake water quality. The multi-parameter water quality tester uses the principle of spectrophotometry to measure water quality parameters. It and a portable temperature and pH meter are used to measure 7 indicators of lake water quality. The correlation coefficient method is used to screen the water quality indicators. Using the model to predict the chemical oxygen demand (COD) of the training data, it shows that the correlation coefficient is above 0.9, and the relative error is controlled within 6%. The research result shows that the fuzzy neural network has high prediction accuracy and generalization ability for water quality indicators with a short prediction time. This is a method that can effectively predict lake water quality.

Using mathematical methods to establish a model to predict water quality indicators can not only obtain high-precision prediction data but also save maintenance and development costs [4].

In order to effectively analyze the lake water quality, this paper uses a multi-parameter water quality meter to measure multiple parameters in the lake. The fuzzy neural network (FNN) was used to predict the COD content in the lake water after collecting multiple sets of water quality parameters. Comparing the predicted values with the actual values, the effectiveness of the FNN in predicting water quality parameters is investigated, which provides a more scientific basis for the water quality evaluation of surface lakes.

Multi-parameter water quality tester(5B-3B(V10); Lianhua Technology Co.) and portable temperature and pH meter are used in this experiment.

In order to effectively predict the content of COD in the lake water, we need to make accurate measurements of the lake water quality. The multi-parameter water quality tester is used to measure five indicators (COD, total phosphorus, total nitrogen, ammonia nitrogen, permanganate index), and portable temperature and pH tester is used to measure the temperature and pH value of the water sample.

In order to decrease the training cost of the FNN prediction model and enhance the prediction accuracy, the Pearson correlation coefficient is used to calculate and rank the correlation between COD and several indicators in this paper. To enhance the convergence speed and operational effect of the FNN, the data is normalized and normalized to [0, 1]. Data normalization processing can greatly increase the training speed of the neural network model and ensure the good performance of the network.

Under normal circumstances, the network model has good predictive performance, and the prediction results are ideal. We time the neural network model when running it, and found that the running time of the network model is controlled within 5 seconds, and the running time is shorter. Therefore, the FNN has the good fitting ability, high prediction accuracy and generalization ability, and the running time is short, which can well predict COD content in the lake water.

Spectrophotometry method is used to measure the various indicators of the lake to obtain the relevant data of the water quality indicator. FNN is applied to predict the COD content by building a prediction model on the basis of total phosphorus, total nitrogen, ammonia nitrogen, permanganate index, temperature, and pH parameters of the lake water.

The training data is predicted, and a high correlation coefficient is found, indicating that the model is high and not overfitting. Using the trained FNN model to predict the COD content of the prediction set data, it is found that the correlation coefficient is almost within 0.8-0.95, and the relative error is controlled within 15%, indicating that the model has good performance, high computing power and short running time, which is suitable for the prediction of complex lake water quality.

PAPER 3: A Tensor Model for Quality Analysis in Industrial Drinking water supply system.

Authors: Di Wu , Hao Wang , Razak Seidu

Drinking Water Supply (DWS) is one of the most critical and sensitive systems to maintain city operations globally. The high standard water quality requirement not only provides convenience for people's daily life but also challenges the risk response time in the systems. Prevalent water quality regulations are relying on periodic parameter tests. This brings the danger in bacteria broadcast within the testing process which can last for 24-48 hours. In order to cope with these problems, we propose a tensor model for water quality assessment. This model consists of three dimensions, including water quality parameters, locations and time. Furthermore, we applied this model to predict water quality changes in the DWS system using a Random Forest algorithm.

This is a three-order tensor model, including Parameter, Location and Time. Parameter order can be divided into concrete water quality parameters, including physical group (*e.g.* Temperature, Conductivity), chemical group (*e.g.* pH, Alkalinity), and biological group (*e.g.* Coliform, Escherichia coli (E.coli)). Location order is separated into cities and towns according to the requirements. For Time order, the present recordings are based on weekly registrations. Therefore, We use *weeks* to show the water quality evolution. An obvious advantage of this model is easily extending for diverse analysis concerns from both research and industry.

In this work, we design an algorithm based on RF to predict water quality parameters. In this algorithm, we define the physical and chemical water quality parameters as input and biological parameters as output. This is based on biological parameters that normally take much longer time to test (e.g. 24–48 hours) and they are direct threats to human's health. Therefore, with this algorithm, we can take the efficient test results of physical and chemical parameters to evaluate harmful bacteria risks. This algorithm takes each city's recordings. Firstly, separate them into training sets and testing sets. Then train the data to generate a regression model. After this model will be tested and evaluated. The results for each city will be evaluated separately.

This model and the corresponding prediction algorithm can generally predict the tendency of biological parameter changes.For the peak value predictions, as the red cross points in the figure, the accuracy is not satisfactory.We have evaluated the general accuracy using RMSE, it ranges from 0.22 to 0.55. Coliform has better prediction accuracy. Clostridium perfringens is the worst. The reason behind needs to be interpreted by domain expertsThe prediction time efficiency is on average 5.26 seconds. This fulfills the requirements of the DWS systems

PAPER 4: Applying Data mining and HPC for water quality assessment and prediction.
Authors: Ruijian Zhang, Hairong Zhao, Yong Piao

Water quality assessment and prediction of Lake Michigan are becoming major challenges in Northwest Indiana. Traditionally, mechanistic simulation models are employed for water quality modeling and prediction. However, given the complicated nature of Lake Michigan in Northwestern Indiana, the detailed simulation model is extremely simple in comparison and, at some point, additional detail exceeds our ability to simulate and predict with reasonable error levels. In this regard, this research applied data mining technologies, as an innovative alternative, to develop an easy and more accurate approach for water quality assessment and prediction. The drawback of data mining modeling is that the execution takes quite a long time, especially when we employ a better accuracy but more time consuming algorithm in clustering. Therefore, we applied the High Performance Computing System of the Northwest Indiana Computational Grid to deal with this problem. The experiments have achieved very promising results. The visualized water quality assessment and prediction obtained from the research are published in an interactive website so that the public and the environmental managers can use the information for their decision making.
The objective of this research is to explore the consideration of improving water quality assessment and prediction by applying data mining technology for more accurate and easier-implemented modeling,
Data mining is the process consisting of the following steps: data collecting and data preparation, clustering, classification and prediction. The framework of this project is to design and analyze the data mining methods for water quality modeling. To conduct the examination of this modeling, we will design and implement an optimal algorithm using an enumerative method for better water quality prediction accuracy. The decision tree classification tool C5 will also be employed in this project . These raw data are incomplete and inconsistent in terms of both time periods and water quality attributes. Some values are missing; some values are invalid. We have to perform data cleaning, data format and normalization, missing or invalid data treatment and other data preprocessing and preparation work. For the purpose of simplicity, we chose only five water quality attributes. They are: water temperature, dissolved oxygen, pH value, specific conductivity and turbidity.
We applied clustering to assign cases of a dataset into subsets (called clusters) so that data in the same cluster are similar in some sense. The similarity can be measured by the distance of each other. In the experiments, the algorithm we designed assigns the water quality data into 3 clusters of quality levels: good, fair and poor.
We designed and implemented an optimal algorithm based on enumeration. The complexity of the algorithm is in the order of n to the power of k (n represents the number of cases in the dataset and k represents the number of clusters). It usually takes quite a long time to execute, but it could guarantee the convergence to the global optimum.
Applying parallel programming and the high performance computing system, the execution time for the enumerative algorithm could be reduced to an acceptable level.
We employed C5, a decision tree based data mining tool, to perform the further classification (assessment) and prediction [1].
In order to improve the performance, we applied parallel programming on the enumerative algorithm. The parallel computing assigns one process as the master process and others as slave process. The master process reads in the data, broadcasts the data and other information to the slave processes. Finally the master processor receives the results from each slave process and chooses the global optimized clustering
Using the produced decision trees to predict unseen cases, the prediction accuracy rate reached 82 percent, two percent better than applying k-means algorithm and about the same improvement as applying the mechanistic simulation models. It is anticipated that if we use more attributes in the proposed project in the future, the accuracy rate could be further improved.

PAPER 5: Data-Driven Water Quality Analysis and Prediction
Gaganjot Kang, Jerry Zeyu Gao, Gang Xie

Water quality becomes one of the important quality factors for the quality of life in smart cities. Recently, water quality has been degraded due to diverse forms of pollution caused by disposal of human wastes, industrial wastes, automobile wastes. The increasing pollution affects water quality and the quality of people's life. Hence, water quality evaluation, monitoring, and prediction become an important and hot research subject. In the past, many environmental researchers have dedicated their research efforts on this subject using conventional approaches. Recently, many researchers began to use the big data analytics approach to studying, evaluating, and predicting water quality due to the advances of big data applications and the availability of environmental sensing networks and sensor data. This paper reviews the published research results relating to water quality evaluation and prediction. Furthermore, the paper also discusses the future research needs and challenges.

Water quality evaluation is an important way to monitor and control water pollution. Water quality evaluation shows how well the quality of water can meet the requirements of the user.t is defined in terms of certain physical, chemical and biological characteristics. These characteristics are traditionally collected manually from different water resources (i.e. lakes, rivers, and oceans), and assessed manually. There are various factors that may affect the quality of water which include thermal pollution, acidification, salinization, ion toxicity etc.the conventional methods for water quality evaluation can be classified into two classes:

- Single factor based methods
- Comprehensive index based methods

In a single factor based method, the most impaired water quality parameter is considered to evaluate water quality.The researchers in [5] have constructed a two-level index system, whose first-class indexes consist of physical indexes, organic matter, heavy metal, nutrients, oils, radioactive material, and new toxic pollutant.

Water Quality Model Processes

- Mass-balance principle: The basic principle of water quality models is that of mass balance.
- Design stream-flows: In streams and rivers, the water quality may vary significantly, depending on the water flow.
- Temperature: Temperature affects almost all water quality processes taking place in water bodies. Thus, it is an important factor while creating a water

Least squares support vector machine model presents a supervised learning technique in which one finds the solution by solving a set of linear equations .the water samples were collected from 100 different places in different time and locations, including urban domestic sewage and surface water, specifically including river, lake, seawater, waste water from car washer, etc.Before the training processing, a data pre-processing is performed.The characteristics of fluorescence data are high-dimensional, spectral overlap, nonlinear, and so on. To solve those problems, they run a blank assay to eliminate the interference of spectral data, such as ambient noise, and temperature drift.After the clustering process, LSSVM algorithm is used to establish the predicting regression model between spectral data and TOC index by TOC-VCSH of training samples for each cluster.After that, the final results can be given for a new sample and its vector angle cosine as a criterion is used to judge the appropriate cluster and regression model in order to get analysis value.

There are lots of sensor data quality issues which affect the accuracy of water quality evaluation and assessments due to device faults, battery issues, and sensor network

communication problems.

As the advance of smart sensing and IoT, more and more environmental sensors (including water sensors and networks) have been installed and deployed for many water resources, such as lakes, rivers, creeks, ocean bays and coasts. However, there is a lack of integrated real-time big data based water evaluation and monitoring environments for smart cities to support dynamic water quality evaluation, monitor, and supervision management. Water in a city could be considered as a multi-level water system, covering surface water and underground water. The water quality on both levels usually affect each other.

With the advance of IoT infrastructures, big data technologies, and machine learning techniques, real-time water quality monitoring and evaluation is desirable for future smart cities. This paper reports our recent literature study, reviews and compares current research work on water quality evaluation based on big data analytics, machine learning models and techniques. Finally, it highlights some observations on future research issues and challenges.

Paper 6: Prediction of water quality using Naive Bayesian algorithm
Authors: P. Varalakshmi; S. Vandhana; S. Vishali

Environmental quality is a degree of the health of environments that includes living organisms and the effects it has on the comfort and psychological state of the people . There are various methods of measuring the quality of the water that relies on the purpose for what the water is used for. Water is basically tested for its physical content, chemical content and biological content.

Raymond RaLonde has discussed the partition triangle method for categorizing water based on pH, organic matter concentration, dissolved solids and sediment load. The classification here is meant for wastewater management and treating industrial water. The model has been tested with Alaskan water measuring temperature, salinity, turbidity, dissolved solids, chloride, pH to identify the nature of the water.

A number of water quality initiatives applied Hazard Analysis and Critical Control Points (HACCP) principles and steps to control the water borne diseases. It presented a way to get used to the HACCP approach to drinking water systems . A.H. Havelaar has proposed the integration of total quality management and quantitative risk assessment to ensure safe drinking water by choosing effective critical control points for different types of water.

Gruttner, Frank and Kristense have proposed a decision tree to recognize and evaluate the critical control points. The methodology can be used in existing HACCP based Quality management systems. The fundamental approach of the method is to apply step-by-step procedure to assess the entire water system. .

Dewan Md. Farid has proposed an adaptive Naive Bayes tree for multi class classification for assessing the water quality. It is a concept proposed on the basis of intelligence decision making .

The existing methods mostly analyze the water with the tests but don't focus on its classification. The proposed model of this paper classifies the water type, with the values obtained periodically over recent years.

The model is prone to few drawbacks, like, it does not mention the factors that makes it unfit for drinking. In the near future, an algorithm for the suggestion generation component (to make the water fit for drinking) of the proposed architecture will be devised.