RESEARCH ARTICLE OPEN ACCESS

WATER POLLUTION THROUGH SVM CLASSIFICATION USING SOFTWARE TECHNIQUES

Dr.C.Brintha Malar*, Dr.K.Siva Sankar**

*(Department of Physics, Udaya School of Engineering, and Vellamodi)

** (Department of Information Technology, Noorul Islam Centre for Higher Education, and Kumaracoil)

--************************----------------------------------

Abstract:

This paper notifies the automated discharge location suggestion system for wastewater treatment. The quality of water is affected when more harmful substances mix up with the water and the water is not fit for consumption. Additionally, the environment is also polluted by the wastewater. The proposed automated discharge location suggestion system is based on four phases such as data pre-processing, attribute selection, statistical feature extraction and classification.

The statistical features such as mean, standard deviation and variance are extracted from the data and it forms the base of the classification task. Multiclass SVM with different kernel functions are utilized for achieving the classification task and this work concludes that RBF SVM performs well than poly SVM and sigmoid SVM.

Keywords **—Classification Waste water discharge, pollution control, supervised learning, water pollution.**

--************************----------------------------------

I. INTRODUCTION

Water is the basic requirement of all the living organisms and life is impossible without water. Researchers want to know, if any other planet possess water to live in. Hence, water has got that much value. Water is not only used for oral consumptionbut also exploited in domestic purposes such as washing, cleaning, cooking and so on.

Additionally, water is the major component of all the things that can be seen, consumed and felt. For instance, plants require water for their growth and so do the animals. However, the fresh water constitutes only three percent of the whole water source in the earth. HinrichsenD. and Tacio H. (2002) states 0.01% of water is accessible by humansout of all the water sources. The most common pollutants of water are bacteria, metals, pesticides and so on.All the common pollutants of water are summarized below.

Bacterial growth is one of the serious problems found in the drinking water, which is very risky for the human health. Natural resources such as rivers and lakesare polluted by bacterial growth, as stated by Aziz J.A. (2005).Most of the portion of drinking water is polluted by bacteria such as coliform and Escherichia coli (E.coli). E.coli may cause diarrhoea in association with stomach pain. The groundwater is also polluted by bacteria, in case of leakage in sewage, septic tanks and other wastages, stated by SharA.H. et.al. (2008).Some of the general bacteria found in drinking water are listed below.

- ❖ Acinetobacter baumannii
- ❖ Aeromonashydrophila
- ❖ Citrobacterfreundii
- ❖ Edwardsiellatarda
- ❖ Klebsiella species

The Acinetobacter baumannii may cause pneumonia and affects the urinary track, bloodstream of the humans, as cited by Gerischer

(2008). Chopra et.al. (2000) claims the Aeromonashydrophilamay affect the tissues of the living organisms.

The bacteria that may cause infection in urinaryand respiratory track, which is claimed by Badger et.al. (1999).Edwardsiellatarda may induce diahorreaand gastroentities, as stated by Verjan et.al. (2005). Podschun and Ullmann (1998) addressed that the bacteria Klebsiella species may cause pneumonia, septicemia and spondylitis. This subsection presents some important information about the contamination of bacteria in water.

II. BACKGROUND

Hamed M.M. et.al. (2004) the performance of a waste water treatment plant is predicted by means of Artificial Neural Networks (ANN). The parameters being used by this work are Biochemical Oxygen Demand (BOD) and Suspended Solids (SS). The relationship between the data is studied by the ANN, which makes it possible to predict the performance of waste water treatment plant.

A three layer ANN is proposed for predicting the Chemical Oxygen Demand Removal Efficiency (CODRE) in the cotton textile waste water. The training process of this work is attained by means of Back Propagation (BP) training in association with Principal Component Analysis (PCA), as proposed by Yetilmezsoy K and Zengin Z.S. (2009). The performance of ANN is studied and compared.

Zeng G.M. et.al. (2003) studied the waste water treatment process of paper mill by means of neural networks. The relationship between the entities of the data is studied by means of multilayer back propagation neural networks. This system has shown better prediction rates.

A hypersaline oily waste water is processed by ANN by means of Feed Forward Neural Network (FFNN), proposed by Pendashteh A.R. et.al. (2011). The neural network is trained by back propagation algorithm and the parameters being used are COD, Total Organic Carbon (TOC) and concentrations of oil and grease.

Sanayei Y. et.al. (2014) proposed a ANN based dye containing waste water treatment technique. This neural network works in association with Wiener-Laguerre model. The parameters being selected by this work are COD, Mixed Liquor Volatile Suspended Solids (MLVSS) and reaction time. The parameters are modified by Levenberg-Marquardt (LM) algorithm.

Jing L. et.al. (2014) states that marine oily wastewater being polluted by polycyclic aromatic hydrocarbon, which is called naphthalene is treated by ANN. The parameters of this system include fluence rate, salinity, temperature, initial concentration and reaction time. The ANN is configured by feed-forward based LM algorithm.

Granata F. et.al. (2017) employ machine learning algorithms such as Support Vector Regression (SVR) and Regression Trees (RT) to predict the quality of water by the indicators. It is shown that SVR performs better than RT.

III. SVM CLASSIFICATION

A. *Introduction*

Water is the most basic and primary need of all living organism. It is an absolute need to preserve the available water. Yet, due to the technological advancement, most of the wastages from the industry are released to water sources, which lead to severe pollution. In order to combat with this issue, the government bodies formed a pollution control board.

The pollution control board sets some limits for the pollutants or the chemical parameter of the wastewater. This work attempts to propose an auto-suggesting discharge location system for wastewater treatment. The discharge locations being considered by this work are inland surface, irrigation land and marine coastal area.

The better discharge locations are suggested by the classifier Support Vector Machine (SVM), which is trained by the statistical features. The performance of the proposed approach is tested against accuracy, sensitivity and specificity rates by varying the kernel functions.

B. *Proposed Automated Water Discharge Location Prediction System*

The proposed supervised learning based approach aims to provide the basic knowledge of the general standards for the water discharge which contains about 31 parameters. The overall flow of the work is depicted in Fig. 1

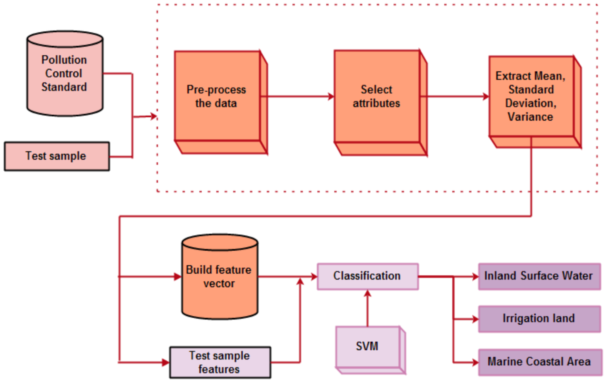

Fig. 2Overall flow of the work

Initially, the input data is pre-processed and normalized. The statistical features are extracted from the pre-processed data. Finally, during the process of classification, the classifier makes a decision about the preferable discharge areas. The detailed explanation of all the phases in the proposed approach is presented as follows.

1) ***Data Pre-processing:***The input database of this work contains nearly thirty one attributes and the range of these attributes is not standard. This data pre-processes the input data by performing the autofill operation. For instance, certain columns in the dataset may not contain values and the pre-processing step attempts to add zero in that place.

For the better execution of any algorithm, it is good to avoid empty fields. Some of the important attributes being considered by this work are Total Suspended Solids (TSS), pH, BOD, temperature, COD, phenolic compounds, Fluoride (F), Sulphide (S), pesticide, detergents and so on.

2) ***Attribute Selection:***The pre-processed data is then passed to the next phase called attribute selection. This phase makes the test sample data to comply with the training data. For instance, the test sample data may not contain all the attributes being present in the train dataset.

Hence, this operation considers only the attributes of the test sample and during the process of comparison, all the attributes that are not the part of the train dataset are modified as zero. By this way, the attribute selection process work and is performed only during the testing process. The overall algorithm of this work is presented as follows.

Algorithm for Auto-suggesting Discharge Location System for Wastewater

```
// Training
Input: General standards from pollution control board
Output : Knowledge gaining
Begin
    Pre-process the general standards by autofill operation;
    For all records
      do
         Extract mean ($M$), standard deviation ($SD$) and Variance ($V$) features;
         Construct $fv(TD)$ and store it in the local database;
         Feed the knowledge to SVM classifier;
      End;
End;
// Testing
Input: Measurement of pollutants in the sample wastewater
Output : Optimal discharge area suggestion
Begin
    Pre-process the pollutant list by autofill operation;
    For the test sample
      do
         Extract $M, SD, V$ features;
         Construct $fv(TD)$;
         Apply SVM classifier to match the test and train samples;
         Utilise $Poly\ SVM, Sigmoid\ SVM$ and $RBF\ SVM$ for matching;
          Analyse the performance;
      End;
End;
```

3) ***Statistical feature extraction:***As soon as the attributes are selected, the statistical features are extracted from the data. Feature extraction is the heart of the classification system and the efficiency of the classification system depends on the potential of the features being utilized. Hence, the features should be more precise and crispy, such that the classification task can be performed effectively.

When the training data is fed into the discharge location classification system, the system extracts the statistical features such as mean, standard deviation and variance of the training data and forms the feature vector. The so formed feature vector is saved in the database for performing future classification tasks.

A short feature vector may be inefficient, as the details of the features are insufficient for performing accurate classification. On the other hand, a large feature set involves time, memory and computational complexity.

Taking this into account, the proposed approach attempts to build an efficient and sharp feature vector that can bring in better accuracy with lesser time complexity. The mean, standard deviation and variance are computed as follows.

$$\mathrm{M} = \frac{1}{n}\sum_{i=1}^{n} D_i \qquad (3.10)$$

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}|D_i - \mathrm{M}|^2} \qquad (3.11)$$

$$V = \frac{\sum D^2}{n} - M^2 \qquad (3.12)$$

All the three features work in a blended fashion and arrives at better feature set and the feature vector is formed by

$$fv(TD) = \{|FV_i(f(M, SD, V)|\}; i = 1,2, \dots, n \qquad (3.13)$$

In the above equations, $D_i$is the input data, TD is the train data. The computed feature vectors are stored in database for future classification problem. When the training process gets completed, the system is ready for performing the testing phase.

In this stage, the known values of the chemical parameters are passed in to the automated discharge location classification system. The proposed work suggests the best possible discharge location after analysing the parameters of the wastewater. The following section presents the SVM classification process.

4) ***SVM Classification:***This work utilizes multiclass SVM for selecting the best possible discharge location of the wastewater after analysing the quality of water. As this work considers three different discharge locations such as irrigation land, inland surface and marine coastal area, multiclass SVM is employed.

SVM can be incorporated in two different ways for dealing with a multiclass classification issue. The first way employs multiple binary SVM classifiers and each classifier is trained to handle a single class out of a set of classes.

In the second way, multiple classifiers are processed over every pair of classes and the data is assigned to the class with maximal votes. This technique exploits n(n-1)/2 classifiers and finally max-vote strategy is followed, as presented by Hsu C.W. et.al. (2002). This work performs multiclass classification by tackling all the classes at the same instant of time by solving a single objective function.

The classification problem with n different classes is denoted by a single optimization problem and is written as

$$\min_{\mathrm{w,b,\omega}} \frac{1}{2} \sum_{y=1}^{n} w_y^p \, w_y + C \sum_{i=1}^{l} \sum_{y \neq s_i} \omega_{i,y} \qquad (3.14)$$

where

$$w_{s_i}^p \rho(x_i) + b_{s_i} \geq w_y^p \rho(x_i) + b_y + 2 - \omega_{i,y};\ \omega_{i,y} \geq 0 \qquad (3.15)$$

where $i = 1,2, \dots l$ are training samples and $y \in \{1,2, \dots n\}$. The final decision is obtained by the below given equation.

$$dec_{fn} = \max_{\mathrm{y=1,2,..n}} (w_y^p \rho(x_i) + b_y) \qquad (3.16)$$

This way of classification conserves more time and is efficient. Besides this, the requirement of support vectors is lesser, when compared to the usage of multiple binary SVMs. Thus, the multiclass SVM can serve its purpose, irrespective of the class count. The classification process is carried out by this way and the performance of the system is analysed as follows.

IV. RESULTS AND DISCUSSION

This work trains the classifier with the standard data by the pollution control board, which is downloaded from http://www.environmentallawsofindia.com/ tolerance-limits-for-trade-effluents.html. This standard contains about thirty one attributes. The quality of the water can be predicted by means of this standard and based on the quality, the water is suggested to get discharged in specific areas.

Based on this standard, the SVM is trained and when a test data sample is passed as input, the SVM pre-processes the data, selects the attribute, extracts the feature and compares the test feature vector with the train feature vector. By this way, the SVM determines the best suitable discharging area by taking the quality of water into account.

The performance of the proposed approach is tested in terms of standard performance measures such as accuracy, sensitivity and specificity and is compared against different kernels of SVM such as poly, RBF and sigmoid.

From the experimental results, it is evident that RBF SVM performs better than the other two. The experimental results of the proposed approach are tabulated in Table I.

TABLE I
EXPERIMENTAL RESULTS OF THE PROPOSED WORK

Performance Metrics	Poly SVM	Sigmoid SVM	RBF SVM
Accuracy (%)	97.1	96.6	98.4
Sensitivity (%)	94.2	95.5	98.1
Specificity (%)	93.6	94.9	96.9
Time consumption (ms)	1832	1807	1742

The above presented Table tabulates the experimental results attained by the proposed approach of this research phase. This work is different from the previous work in terms of feature extraction and the employment of classifier. The previous work utilizes mean and standard deviation, whereas this work utilizes mean, standard deviation and variance. Additionally, the performance of SVM classifier is better than the k-NN classifier.

The experimental results prove that the performance of this work is better than the existing approach in terms of accuracy, sensitivity and specificity. However, the time consumption of the work is little bit greater than the existing approach. The graphical representations of the results are presented from Fig.2 to Fig. 5.

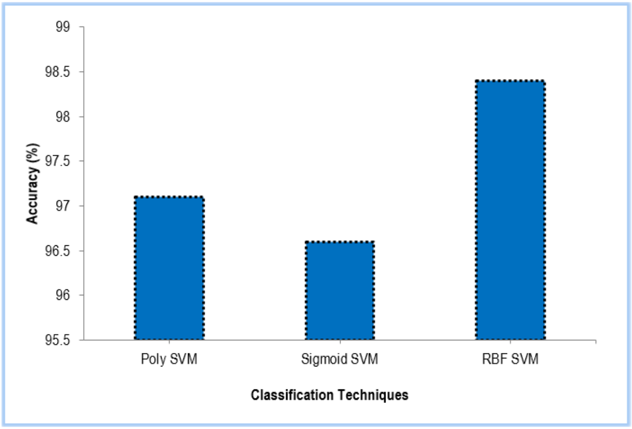

Fig. 3Accuracy rate analysis

The accuracy rate of the proposed approach is analysed by varying the kernels of SVM. The performance of three kernel functions such as poly SVM, sigmoid SVM and RBF SVM are analysed. The accuracy rate of a classification algorithm is so important, as the incorrect classification reduces the reliability of the classifier.

From the experimental analysis, it is evident that the accuracy rates of RBF SVM is 98.4 percent and is the highest accuracy rate. The poly SVM shows the least accuracy rates, which is 97.1 percent. The following Fig. 3presents the sensitivity rates of the proposed approach.

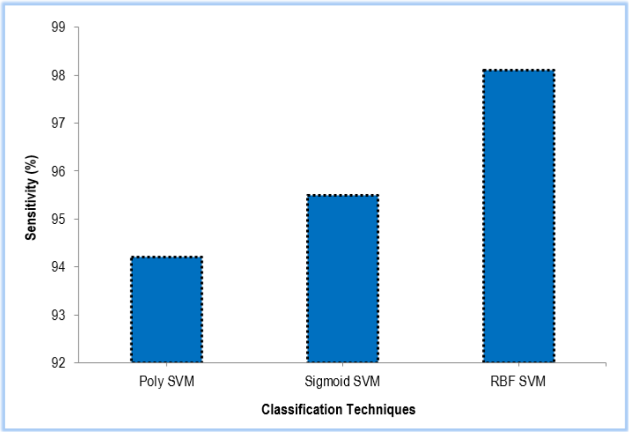

Fig. 3 Sensitivity rate analysis

The sensitivity rates of the proposed approach is analysed and the results are presented in graphical format. The sensitivity rate of RBF SVM is better than the other two SVM kernels. The RBF SVM shows the sensitivity rate of 98.1 percent.

The poly and sigmoid SVM shows 94.2 and 95.5 percent sensitivity rates respectively. This makes sense that the false negative rate of RBF SVM is lesser and this boosts up the sensitivity rates. The following Fig. 4 presents the specificity rates analysis of the proposed approach.

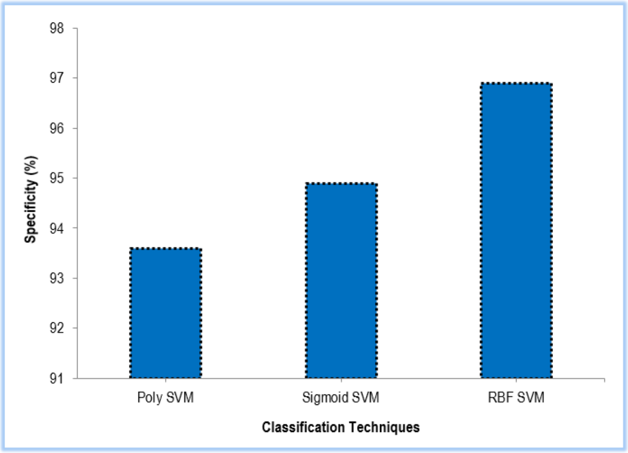

Fig. 4Specificity rate analysis

The specificity rate of the proposed approach is depicted in the above presented graph. The specificity rates denote that the classification system involves lower false positive rates. False positive rates mean that the classification entity actually do not belong to class ‘A’ but belong to class ‘B’, yet the entity is classified to be an entity of class ‘A’.

The specificity rate proven by the RBF SVM is 96.9 percent and is the greatest. The second better performing classification technique is poly SVM, which proves 93.6 percent. The following Fig. 5 presents the time consumption analysis of the proposed approach.

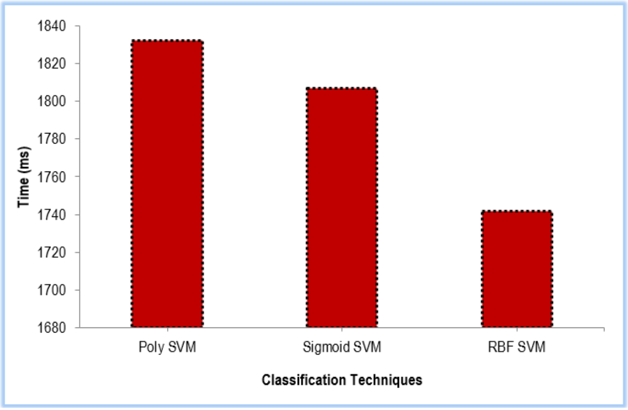

Fig. 5Time consumption analysis

The time consumption of the proposed approach is compared by varying the kernel functions of SVM. Though the time consumption of all the three kernel functions of SVM are more or less similar to each other, RBF SVM consumes 1742 ms, which is the least and the remaining classification techniques consume 1832 and 1807 ms respectively.

On analyzing the performance of different kernels of SVM, it is found that the RBF kernel of SVM serves well than the other two. The greatest accuracy, sensitivity and specificity values of SVM are proven by RBF SVM. Hence, this work concludes that the performance of RBF SVM is satisfactory in terms of better accuracy, sensitivity and specificity rates.

V. CONCLUSIONS

This phase presents an automated discharge location suggestion system for wastewater treatment. The quality of water is affected when more harmful substances mix up with the water and the water is not fit for consumption. Additionally, the environment is also polluted by the wastewater. The proposed automated discharge location suggestion system is based on four phases such as data pre-processing, attribute selection, statistical feature extraction and classification.

The statistical features such as mean, standard deviation and variance are extracted from the data and it forms the base of the classification task. Multiclass SVM with different kernel functions are utilized for achieving the classification task and this work concludes that RBF SVM performs well than poly SVM and sigmoid SVM.

REFERENCES

[1] Maged M Hamed, Mona G Khalafallah, Ezzat A Hassanien, "Prediction of wastewater treatment plant performance using artificial neural networks", *Environmental Modelling & Software*, Vol.19, No.10, pp: 919-928, 2004.

[2] KaanYetilmezsoy, ZehraSapci-Zengin, "Stochastic modeling applications for the prediction of COD removal efficiency of UASB reactors treating diluted real cotton textile wastewater", *Stochastic Environmental Research and Risk Assessment*, Vol.23, No.1, pp.13-26, 2009.

[3] G.M.Zeng, X.S.Qin, L.He, G.H.Huang, H.L.Liu, Y.P.Lin, "A neural network predictive control system for paper mill wastewater treatment", *Engineering Applications of Artificial Intelligence*, Vol.16, No.2, pp.121-129, 2003.

[4] Ali Reza Pendashteh, A. Fakhru'l-Razi, NazChaibakhsh, LuqmanChuahAbdullah,SayedSiavashMadaeni, ZurinaZainalAbidin, "Modeling of membrane bioreactor treating hypersaline oily wastewater by artificial neural network", *Journal of Hazardous Materials*, Vol.192, pp.568-575, 2011.

[5] YasamanSanayei, NazChaibakhsh, Ali Chaibakhsh, Ali Reza Pendashteh, Norli Ismail,Tjoon Tow Teng, "Long-Term Prediction of Biological Wastewater Treatment Process Behavior via Wiener-Laguerre Network Model", *International Journal of ChemicalEngineering*,Vol. 2014, pp. 1-7, 2014.

[6] Liang Jing, Bing Chen, Baiyu Zhang, "Modeling of UV-Induced Photodegradation of Naphthalene in Marine Oily Wastewater by Artificial Neural Networks", *Water, Air, & Soil Pollution*, vol.225, 2014.

[7] Francesco Granata 1, Stefano Papirio, Giovanni Esposito, Rudy Gargano, Giovanni de Marinis, "Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators", *Water*, Vol.9, No. 105, pp. 1-12, 2017.

[8] Taufiqurrahman ;Ni'am Tamami ; DitoAdhi Putra ; Tri Harsono, "Smart sensor device for detection of water quality as anticipation of disaster environment pollution", International Electronics Symposium, Denpasar, Indonesia, 29-30 Sep, 2016.

[9] XuLuo; Jun Yang; Li Chai, "Water pollution source detection in wireless sensor networks", *IEEE International Conference on Information and Automation*, Lijiang, China, 8-10 Aug, 2015.

[10] AsmaaHashemSweidan ; Nashwa El-Bendary ; Aboul Ella Hassanien ; Osman Mohammed Hegazy; Abd El-karim Mohamed, "Machine Learning based Approach for Water pollution detection via fish liver microscopic images analysis", *International Conference on Computer Engineering & Systems*, Cairo, Egypt, 22-23 Dec, 2014.

[11] M. Soprani; O. Korostynska; A. Mason et.al, "Low-Frequency Capacitive Sensing for Environmental Monitoring of Water Pollution with Residual Antibiotics", *International Conference on Developments in eSystems Engineering*, Liverpool, UK, 31 Aug-2 Sep, 2016.

[12] XuLuo; Jun Yang, "Problems and challenges in water pollution monitoring and water pollution source localization using sensor networks", *Chinese Automation Congress*, Jinan, China, 20-22 Oct, 2017.

[13] Salah Farhan A. Sharif; Ghada Mahdi Al-Saadi, "Evaluation of air and water pollution caused by South Baghdad Power Plant South Baghdad Power Plant", *International Conference on Environmental Impacts of the Oil and Gas Industries: Kurdistan Region of Iraq as a Case Study (EIOGI)*, Koya-Erbil, Iraq, 17-19 April, 2017.

[14] ZhengFeng; Jun Yang; XuLuo, "A water pollution source localization method in three-dimensional space using sensor networks", *IEEE Conference on Industrial Electronics and Applications*, Siem Reap, Cambodia, 18-20 June, 2017.

[15] C. Brintha Malar, Dr. S. Akilandeswari, "Optimal Area Prediction for Waste Water Discharge by Employing Supervised Learning" *International Journal of Current Advanced Research*, Vol.6, No.12, pp.8523-8526, 2017.

[16] C. Brintha Malar, Dr. S. Akilandeswari, "Auto-Suggesting Discharge Location System for Wastewater Treatment", *International Journal of Current Research and Modern Education*, Volume 3, Issue 1, Page Number 213-217, 2018.