**Predicting Water Quality Index for farm irrigation**

**Abstract**

Agriculture sector of Indian economy in which more than half of the population is involved, contributes to less than quarter of the GDP. With advancement in ICT, tools and techniques can be developed that can help in analyzing and automating various phases of farming. This paper focuses on prediction of a water quality index which indicates the degree of sodicity and salinity of water. Five parameters of water: Sodium Absorption Ratio(SAR), Electrical Conductivity(EC) and concentration of Sodium, Chloride and Bicarbonate ion are measured for water samples using which the Irrigation Water Quality Index(IWQI) is calculated. Five parameters are further reduced to three parameters using correlation analysis and a classification model is developed using seven classification techniques for prediction of IWQI class. Best result is given by Random Forest Classifier(86.9%) followed by Gradient Boosting and Neural Network Classifier. The model developed in the paper can be used in agriculture to save time and cost of expensive lab tests for irrigation water.

**Introduction**

In a survey conducted in 2018, it was found that 50% of the workforce in India is involved in the agriculture sector, however its contribution is only 16% to the GDP. This trend can be seen all over the world, the major reason for which being wastage of resources used for farming. There exists a need to increase the efficiency of each stage in farming and at a cheap cost so that it is affordable. With improvement in technology, these needs can be addressed using innovative solutions like the Internet of Things(IoT). This technology allows us to use sensors which in turn can be used to obtain large datasets for physical and chemical characteristics of soil, water and weather. This data can help in proper utilization of resources and increase the overall efficiency of crop production.

In this paper, we develop a model for prediction of Irrigation Water Quality Index(IWQI) that mainly indicates the degree of sodicity and salinity of water sample. We used seven classification techniques which are Support Vector Classifier, Neural Networks, Gradient Boosting , Random Forest, Decision Tree, Bagging and Naive Bayes classifier.

The rest of the paper is organized as follows: section 2 describes the related works, section 3 discusses the preliminary topics, section 4 describes the database and algorithm of the classification model, section 5 gives the results and finally section 6 concludes the paper.

**RELATED WORKS**

With advancement in ICT, many new tools and techniques have been developed for agriculture sector. The term Precision agriculture was developed in which various types of data such as time series, geolocation, sensor etc. are gathered and analysed to improve the efficiency of agricultural practices. For example, crop color and size can be measured using optical sensors and models can be developed for predicting optimal time of harvesting based on images. Irrigation is one of the important phases of agriculture which involves adding water to soil after plantation which needs proper analysis of chemicals in water and quality of water for proper plant growth. Water quality Indices(WQIs) are one way to aggregate various parameters of water samples into a single value. Orozco et al aggregated chemical properties of water and developed a water quality index(WQI) for drinking water based on WHO guidelines. Lerios et al developed a prediction model for prediction of WQI for drinking water using parameters like acidity and temperature of water. The WQI calculated for drinking water was not suitable for use in irrigation water. Mark et al. studied the effect of saline water on plants and proposed methods for using them for irrigation while Shrivastav et al. studied the effect of salinity on soil nutrients which decreases with increase in salinity due to its ability to reduce soil permeability. Meireles et al developed a WQI for irrigation water (which mainly focused on salinity and sodicity of water) using factor and principal component analysis to reduce thirteen parameters of water samples to just five parameters. Vij et al proposes the use of wireless sensor network to measure soil parameters such as moisture, temperature, weather etc. to analyse and predict the amount of water required for irrigation.

**Preliminaries**

**Development of Irrigation Water Quality Index**

Calculation of water quality for irrigation is necessary to study its effect on plant and soil health. Water Quality Index (WQI) is used for representing the overall quality of water. WQI represents the aggregate of various parameter (chemical and physical) values of water and is calculated by multiplying relative weights of parameters with parameter value. The parameters chosen should be able to be measured in all the water sources used for irrigation. In this work, we use the Irrigation Water Quality Index (IWQI) based on salinity and sodicity of water. Next subsections describe the water quality parameters, their weights and formula for calculating IWQI.

**Parameters for IWQI**

The IWQI used in this work was formulated by Meireles et al who reduced 13 parameters of water using Principal Component Analysis(PCA) and Factor Analysis(FA) to 5 parameters which are: Sodium, Chloride and Bicarbonate ion concentration, Electrical Conductivity(EC) and Sodium Absorption Ratio (SAR). Electrical conductivity is generally measured by passing current through solutions and the value of SAR is given by the combination of Calcium, Magnesium and Sodium ion concentration whose formula is given by Equation (). It is the primary indicator of sodicity of water which affects the permeability of soil and water infiltration rate. Both cases of water infiltration rate being too high or too low affects the growth of plants. The weights of the 5 parameters, which represents their contribution to water quality are converted to relative weights by summing up all weights and dividing the each weight by the result. The relative weights are mentioned in table ().

**Quality Measurement Values and IWQI**

The value of parameters measured for water sample should neither be too high not too low for optimal growth of plants due to which the parameters are normalized(between 0 to 100) to quality measurement values (qi) according to limits shown in table () and Equation (). The qi values obtained are multiplied with relative weights of their respective parameters and finally added to obtain IWQI whose value is between 0 and 100. IWQI range is divided into various classes. Each class represents the salinity characteristic where water with IWQI in class range 85-100  is suitable for all types of soil and plants, range 70-85 is suitable for soil with moderate permeability and should be avoided in plants with low salt tolerance, range 55-70 is suitable for soil with moderate to high permeability and plants with medium salt tolerance, range 40-55 is suitable for soil with high permeability and plants with high salt tolerance, while water with IWQI below 40 is not suitable for irrigation and they should be properly treated before use. Complete information is given in table ().

**Classification Algorithms**

**ANN**

In ANN, we have an input layer, hidden layers and output layer each consisting of neurons in it. Neurons receive some input and give an output after applying an activation function. When we give an input to this network of neurons, it produces some output which is tested with the known output for that sample. The error is back propagated to the network so that it can adjust the weights for each neuron to improve the network for upcoming samples.

**SVM**

This classification algorithm is very useful as it tries to maximize the margin between any two classes which makes the decision hyperplane more accurate. This algorithm is also able to classify datasets with higher dimensions using the kernel trick which doesn’t make use of vector calculations so that it can save computation time.

**Gradient Boost Classifier**

It is a classifier which uses loss function, weak learners and an additive component to reduce the loss. It has a necessary condition that loss function should be differentiable. This algorithm uses decision trees as its weak learners. It is quite powerful but is prone to overfitting.

**Random Forest Classifier**

It is an ensemble learning method which uses multiple trees. This is beneficial because one classifier may be prone to overfitting but when using a group of uncorrelated classifiers, the result is a refined classifier which doesn’t overfit the dataset. Here each classifier gives its own output and the output which is chosen by more number of classifiers becomes the final output.

**Decision Tree**

It is a classifier which uses a tree-like structure where at each node a decision is made according to the feature in the dataset and final output is given by the leaf node.We can obtain the rules for any output by traversing from root to leaf. This algorithm is simple and easy to implement and also functions as weak learners in Gradient boosting classifier, Bagging classifier and Random forest.

**Naive Bayes**

It is a simple classifier which uses Bayes theorem with an exception that features do not depend on each other. This allows us to easily implement and use the algorithm for classification but its results are not reliable.

**Bagging classifier**

This algorithm uses weak learners like decision trees on a random portion of its dataset, using the results from those trees to produce a final result. This algorithm takes advantage of the fact that the construction process is randomized which helps to improve results of unstable weak learners such as neural networks and decision trees.

**4 Materials and methods**

**Major ions dataset**

The dataset used in our work was obtained from US Geological Survey of Brakwish Groundwater(). It consists of 66 thousand rows and contains values for concentration of dissolve solids, metals, nutrients, ions and physical properties such as conductivity out of which only EC, Cl, Na, HCO3, Ca and Mg are used for IWQI. EC stands for electrical conductivity which is measured in milliSiemens/cm. Cl, Na, HCO3, Ca and Mg represents  Chloride, Sodium, Bicarbonate, Calcium and Magnesium ion concentration respectively. These concentrations were measured in mg/L which we converted to millimoles/L. EC needs to be in dS/cm so it was divided by 100. We use Ca, Mg and Na and calculate SAR using Equation 1.

**Irrigation Water Quality Index based classification**

In this work, we develop a classification model and apply it on major ions dataset. This dataset consists of all the necessary features required to obtain the IWQI described in section () and the model is built using these three steps: converting given values to quality measurement values, calculating irrigation water quality index(IWQI) using calculated values and relative weights of each parameter, assigning to a class according to ranges given in Table 2 and finally using different classification techniques and choosing the best of them.

**Quality Measurement Values**

The dataset consists of parameters with different units. Also, the value of any parameter being too low or too high will decrease the quality of water. Hence, we need to normalize the values according to predefined limits of parameters. This is done by converting values of parameters to quality measurement values (qi) that can be obtained using Equation 2 and Table 3. The quality measurement value (qi) is a dimensionless quantity which can then be combined with other parameters without worrying about units. The Table 5 indicates the number of values present in specific ranges of qi for the dataset used.

**CALCULATING IWQI AND DEVELOPING CLASSIFICATION MODEL**

Using the quality measurement values calculated in section (), we multiply them by their relative weights mentioned in table() and add them to obtain IWQI according to eq (). The samples are further classified into quality classes according to quality ranges provided in table (). Correlation matrix of quality parameters which was obtained using Pearson’s Correlation is given in table (). The following observations are made:

* Cl- is highly correlated with only EC and Na
* Na is highly correlated with all the parameters
* EC is highly correlated with all the parameters except SAR
* HCO3 is highly correlated with EC
* IWQI is highly correlated with Na, Cl and EC

 The three parameters: Na, Cl and EC are chosen to develop classification algorithms to predict water quality. Choosing three parameters saves us the cost of measuring HCO3 and SAR. The classification algorithms mentioned in section () are used to predict quality classes and the method having best accuracy and F2 score is selected for further use. The overall algorithm is mentioned in ().

**Evaluation metrics**

For testing our classifier, we use accuracy, precision, recall and F1 score. These metrics are calculated to test the reliability and robustness of classifiers used and are shown for each classifier in Table (). Accuracy helps us to know how much data can our classifier predict correctly. Precision indicates the reliability of a positive output and precision tell us about how much data is correctly classified for a given output. F1 score balances the result of precision and recall, providing us the ultimate metric to evaluate our classifier.

[----------------------------------------------Add equations-------------------------------------------]

**Results of IWQI Classification**

Algorithm 1 is applied to major ions dataset with Electrical Conductivity, Cl and Na as features and IWQI class as labels and results are evaluated using the metrics mentioned in section (), details of which are mentioned in section (). For the major ions dataset, the number of water samples in IWQI range 85-100 is around 1.3% of the total samples which affected the results of correctly classifying samples belonging to that range. This however does not adversely affects the overall result. Random Forest performed the best with an accuracy of 86.9% followed by Gradient Boosting and Neural Networks with accuracies of 85.8% and 84.6% respectively. Naïve Bayes classifier performed the worst with accuracy of 52.6%. The results of all the algorithms are given in table (). We also performed five-fold cross validation and computed the accuracies whose results are in table (). From both the results, performances of all the algorithms are good except that of Naïve Bayes.

**Conclusion**

Analysis of irrigation water quality can help in increasing agricultural productivity and prevent damage to plants. In this work, we developed a classification model for prediction of Irrigation Water Quality Index (IWQI) class. The IWQI used here is an aggregate of five parameters that are SAR, EC, Chloride, Bicarbonate and Sodium ion concentration respectively. For classification, we used seven classification algorithms and selected three out of five parameters: EC, Cl and Na for predicting IWQI class. Random Forest Classifier performed the best followed by Gradient Boosting and Neural Networks. The water quality index used here is mainly representation of salinity of water and the dataset used here is of groundwater. Other parameters such as acidity, oxygen demand etc. can be used in addition to the parameters used here to develop an index covering more factors. Also, dataset covering different types of water bodies can be used to improve the model. This work can be incorporated in precision agricultural systems which can help in saving cost of lab tests for water quality.