Journal of Hydrology 511 (2014) 254–266

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Evaluating influences of seasonal variations and anthropogenic activities on alluvial groundwater hydrochemistry using ensemble learning approaches

Kunwar P. Singh [a,b,*], Shikha Gupta [a,b], Dinesh Mohan [c]

[a] Academy of Scientific and Innovative Research, Anusandhan Bhawan, Rafi Marg, New Delhi 110 001, India
[b] Environmental Chemistry Division, CSIR-Indian Institute of Toxicology Research, Council of Scientific & Industrial Research, Post Box 80, Mahatma Gandhi Marg, Lucknow 226 001, India
[c] School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110 067, India

ARTICLE INFO

Article history:
Received 6 August 2013
Received in revised form 2 January 2014
Accepted 3 January 2014
Available online 27 January 2014
This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Barbara Mahler, Associate Editor

Keywords:
Ensemble learning
Decision tree forest
Decision treeboost
Groundwater hydrochemistry
Seasonal variations
Anthropogenic activity

SUMMARY

Chemical composition and hydrochemistry of groundwater is influenced by the seasonal variations and anthropogenic activities in a region. Understanding of such influences and responsible factors is vital for the effective management of groundwater. In this study, ensemble learning based classification and regression models are constructed and applied to the groundwater hydrochemistry data of Unnao and Ghaziabad regions of northern India. Accordingly, single decision tree (SDT), decision tree forest (DTF), and decision treeboost (DTB) models were constructed. Predictive and generalization abilities of the proposed models were investigated using several statistical parameters and compared with the support vector machines (SVM) method. The DT and SVM models discriminated the groundwater in shallow and deep aquifers, industrial and non-industrial areas, and pre- and post-monsoon seasons rendering misclassification rate (MR) between 1.52–14.92% (SDT); 0.91–6.52% (DTF); 0.61–5.27% (DTB), and 1.52–11.69% (SVM), respectively. The respective regression models yielded a correlation between measured and predicted values of COD and root mean squared error of 0.874, 0.66 (SDT); 0.952, 0.48 (DTF); 0.943, 0.52 (DTB); and 0.785, 0.85 (SVR) in complete data array of Ghaziabad. The DTF and DTB models outperformed the SVM both in classification and regression. It may be noted that incorporation of the bagging and stochastic gradient boosting algorithms in DTF and DTB models, respectively resulted in their enhanced predictive ability. The proposed ensemble models successfully delineated the influences of seasonal variations and anthropogenic activities on groundwater hydrochemistry and can be used as effective tools for forecasting the chemical composition of groundwater for its management.

1. Introduction

Groundwater contamination is a serious global issue today. Continuously increasing level of contamination with a variety of toxic substances and lowering down of the groundwater table due to over-exploitation to meet globally increasing water demand followed by the declining annual recharge have brought them under severe constrains worldwide. Interferences altering the natural water balance have further influenced the redox chemistry of the aquifers resulting in mobilization of several chemical constituents present in the solid matrices (Singh et al., 2007). The chemical composition and hydrochemistry of groundwater in a region are largely determined by the prevalent natural (atmospheric depositions, precipitation, evapo-transpiration, soil/rock-water interactions) and anthropogenic activities (Singh et al., 2005). Since, frequency of occurrence and magnitude of the natural processes and anthropogenic activities in a region vary in time and space; their influences are reflected in the groundwater hydrochemistry, exhibiting wide spatial and temporal fluctuations (Singh et al., 2007). Groundwater resources in the alluvial regions are relatively more prone to contamination due to higher population densities and consequently intense agriculture and industrial activities in these areas (EPA, 1993). Knowledge and understanding of the factors responsible for influencing the groundwater composition and hydrochemistry in a region is essentially required to develop

* Corresponding author at: Environmental Chemistry Division, CSIR-Indian Institute of Toxicology Research, Council of Scientific & Industrial Research, Post Box 80, Mahatma Gandhi Marg, Lucknow 226 001, India. Tel.: +91 522 2476091; fax: +91 522 2628227.
E-mail addresses: kpsingh_52@yahoo.com, kunwarpsingh@gmail.com (K.P. Singh).

http://dx.doi.org/10.1016/j.jhydrol.2014.01.004
0022-1694/

appropriate management strategies for their efficient and timely implementation for the protection of the groundwater resources. Both the short and long term impacts of seasonal variations (monsoon and non-monsoon periods) and industrial activities on groundwater quality in different parts of the globe have been reported (Singh et al., 2007; Mondal et al., 2009). Precipitation during monsoon months may mobilize several surface and sub-surface soil contaminants to the shallow and deeper aquifers, whereas concentration effect may influence the groundwater composition during the non-monsoon periods. Moreover, in non-monsoon periods, several chemical constituents present in the dry aquifer matrix may get oxidized which could be freely available for mobilization to the deeper aquifers during high water table periods (Kumar and Ahmed, 2003). Therefore, to ensure safety of the groundwater resources, it is essentially needed to develop appropriate methods which could be capable of delineating the aquifer hydrochemistry under seasonal and industrial influences with their quantifiable impacts.

Several chemometric modeling methods are now available for predictive purposes, which could offer solution to such a problem. In recent years, the Hasse diagram technique has emerged as an effective tool for water quality assessment and has been successfully applied to different aquatic systems (Voyslavov et al., 2013; Tsakovski and Simeonov, 2014). However, Hasse diagrams are difficult to use for large systems (Elmqvist and Tsigas, 2004). Support vector machines (SVMs) exhibiting excellent generalization abilities with non-linear systems have successfully been used in various research areas (Pan et al., 2008; Singh et al., 2011, 2013), however, these make use of limited data points in model building. In recent years ensemble machine learning methods (Anctil and Lauzon, 2004; Snelder et al., 2009) have emerged as unbiased tools for modeling the complex relationships between set of independent and dependent variables and have been applied successfully in various research areas. In general, these methods are designed to overcome problems with weak predictors (Hancock et al., 2005). Ensemble techniques have the advantage to alleviate the small sample size problem by averaging and incorporating over multiple classification models to reduce the potential for over-fitting the training data (Dietterich, 2000a). Artificial neural networks (ANNs) and decision trees (DTs) are commonly used as base predictors in building ensemble machine learning models (Zhang et al., 2008). However, ANNs suffer from risk of over-fitting in training process (Singh et al., 2011). DTs supplemented with bagging and stochastic gradient boosting techniques improve the prediction accuracy of weak learners (Breiman, 1996). The bagging aims minimizing of prediction variance by generating bootstrapped replica data sets, whereas, boosting creates a linear combination out of many models, where each new model is dependent on the preceding model (Friedman, 2002). Decision tree forest (DTF) and decision treeboost (DTB) implementing bagging and boosting techniques, respectively are relatively new methods for improving the accuracy of a predictive function (Dietterich, 2000b). These techniques are inherently non-parametric statistical methods and make no assumption regarding the underlying distribution of the values of predictor variables and can handle numerical data that are highly skewed or multi-model in nature (Mahjoobi and Etemad-Shahidi, 2008). Here, we have considered the ensemble learning approaches (classification and regression) to investigate the chemical composition and hydrochemistry of the groundwater from two different alluvial regions (Unnao and Ghaziabad) in northern India to evaluate the influences of seasonal variations and anthropogenic activities; and to predict the groundwater contamination using the chemical oxygen demand (COD) as the indicator variable. COD is a synthetic indicator that represents the degree of organic pollution in water (Wu et al., 2011). It is defined as the amount of oxygen equivalents consumed in oxidizing the organic compounds of samples by strong oxidizing agents.

Present research focuses on construction of ensemble machine learning (EML) based classification and regression models for evaluating the impacts of anthropogenic activities and seasonal variations on the composition and hydrochemistry in the selected study areas. Accordingly, the classification models were developed and used to discriminate the groundwater composition under industrial and seasonal influences, identifying the responsible factors. Regression models were developed and used to predict the levels of COD in groundwater samples using set of independent hydrochemical variables. Performances of these models were evaluated in terms of several statistical criteria parameters and compared with the SVM approach as benchmark. This study has shown that the application of EML methods can be useful in predicting the groundwater quality successfully for its effective management.

2. Materials and methods

The basic aim of this study is to find the most accurate possible classification function $\bar{f}_c$ capable of discriminating between the industrial and non-industrial; pre-monsoon and post-monsoon; shallow and deep aquifers groundwater to enumerate the influences of seasonal variations and anthropogenic activities on the groundwater hydrochemistry; and regression function $\bar{f}_r$ capable of predicting the COD levels using the training data pertaining to the groundwater hydrochemistry employing set of hydro-chemical variables measured in the groundwater of the selected study areas using the EML modeling approaches.

Accordingly, in this study, we constructed the DT models for classification and regression. A conventional well known machine learning method, SVM is employed as the benchmark model. Subsequently, the bagging (DTF) and stochastic gradient boosting (DTB) algorithms are incorporated in building tree based ensemble learning models to optimize the prediction accuracy of single decision tree (SDT) model.

2.1. Study area

Here, we have considered two different hydro-chemical datasets pertaining to the groundwater of Ghaziabad and Unnao regions in northern India located in the Indo-Gangetic alluvium plains (Fig. 1). The study region of Ghaziabad covering an area of about 1000 km^2 is located in the Ghaziabad district (28°60′ and 28°76′N latitude; 77°30′ and 77°50′E longitude). The rate of increase in population during last decade was 40.66% and total population of city in 2011 is 1.636 million (Census of India, 2011). Ghaziabad forms physical and hydrological boundary with Delhi, the Capital of India and the Yamuna River, respectively. Ghaziabad is a prominent industrial hub and one of the fastest growing cities in the state of Uttar Pradesh, India. The city houses various types of highly polluting industries including textiles, metal processing industries such as induction and foundries, lead reprocessing units, electroplating or galvanizing and others (pharmaceuticals, chemicals, tanneries, ceramics, pesticides formulations, sugar, food and beverages, distillery, dyes, fertilizers, etc.(Chabukdhara and Nema, 2013). The climate of this region is tropical and has three well demarcated seasons of winter (October–February), summer (March–June), and monsoon (July–September) and receives an average annual rainfall of about 702 mm. The mean temperature in the region varies between 3 to 43 °C. The groundwater table in the region varies between 14 and 25 m below ground level and accessibility to the groundwater is through hand pumps, tube wells, and bore wells. Another data set collected from literature (Singh et al., 2005, 2007) pertains to the groundwater hydrochemistry