Sensors and Actuators B 202 (2014) 279–285

Contents lists available at ScienceDirect

Sensors and Actuators B: Chemical

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

Quantifying mixtures of hydrocarbons dissolved in water with a partially selective sensor array using random forests analysis

James S. Cooper[a,*], Harri Kiiveri[b], Edith Chow[a], Lee. J. Hubble[a], Melissa S. Webster[a], Karl-H. Müller[a], Burkhard Raguse[a], Lech Wieczorek[a,**]

[a] CSIRO Materials Science and Engineering, PO Box 218, Lindfield, NSW 2070, Australia
[b] CSIRO Computational Informatics, 65 Brockway Road, Floreat, WA 6014, Australia

ARTICLE INFO

Article history:
Received 17 March 2014
Received in revised form 8 May 2014
Accepted 21 May 2014
Available online 29 May 2014

Keywords:
Chemometrics
Random forests
Quantification
Chemical sensor
BTEX

ABSTRACT

Mixtures of benzene, toluene, ethylbenzene, *p*-xylene and naphthalene dissolved in water were probed with an array of partially selective gold nanoparticle chemiresistor sensors. A full factorial experimental design was followed to generate every possible combination (unary, binary, ternary, quaternary and quinary). The nominal concentrations of the individual components in the mixtures were 0, 0.5, 1, 5 or 10 mg/L and the combined concentrations were between 0 and 45 mg/L, which are relevant to EPA defined maximum contaminant levels in drinking water. Several different statistical techniques were used to predict the component concentrations in the mixtures based on the sensor array responses. The most accurate technique was a non-linear ensemble method called random forests. The overall root mean square error between the predicted and measured concentrations (residuals) was 0.2–1.5 mg/L for the mixtures with a nominal component concentration of 10 mg/L. The accuracy of the random forests predictions was not unduly affected by increasing mixture complexity. Random forests analysis is a statistical technique suitable for quantifying the relationship between responses of partially selective sensors to the concentration of different hydrocarbons in water.

1. Introduction

Sensor arrays are being more widely adopted in many applications: such as the quality control of food [1], the detection of explosives or contraband [2], and even the diagnosis of medical conditions [3]. These sensors have been made possible through research that has improved the sensitivity, selectivity and robustness of sensing materials [4,5]. The development of sensors has also required improvements to the methods used to analyze the sensor signals [6]. New statistical analysis techniques have been developed and applied to the output from sensor devices to improve the interpolation of their measurements [7].

The complexity of the sensor output and the extent that it will be interpreted will define the best statistical method to use. In particular, sensors that are only partially selective require more advanced statistical methods to be correctly interpreted than perfectly selective sensors. The wrong analysis or a poor understanding of data could easily lead to erroneous classification or prediction [8]. The data analysis of sensor arrays has been thoroughly reviewed [6,9–11], covering many different methods: principal component analysis (PCA) [12,13], principal component regression (PCR) [14,15], extended disjoint principal components regression (EDPCR) [16,17], discriminant analysis (DA) [18,19], nearest shrunken centroid (NSC) [7], k-nearest neighbor (KNN) [20], partial least squares (PLS) [21–23], non-negative least-squares (NNLS) [17], multivariate curve resolution (MCR) [24], multiple linear regression (MLR) [12,23], visual empirical region of influence (VERI) [25], advanced neural networks (ANN) [26–29], back propagated neural networks (BPNN) [30], relevance vector machines (RVM) [31], and support vector machines (SVM) [32]. Each method has its advantages and disadvantages that would make it suitable for different sensor outputs or applications. Surprisingly, a relatively new and elegant technique, random forests, has only had limited application to sensor array data [7,30,33].

The nonlinear multiple regression version of random forests is an ensemble method that uses bootstrap samples and random feature selection to build a large set of regression trees which are averaged to produce predictions [34–36]. The model is built from

* Corresponding author. Tel.: +61 294137143.
** Corresponding author. Tel.: +61 294137975.
E-mail addresses: james.cooper@csiro.au (J.S. Cooper), Harri.Kiiveri@csiro.au (H. Kiiveri), Edith.Chow@csiro.au (E. Chow), Lee.Hubble@csiro.au (Lee.J. Hubble), Melissa.Webster@csiro.au (M.S. Webster), Karl.Muller@csiro.au (K.-H. Müller), Burkhard.Raguse@csiro.au (B. Raguse), lech.wieczorek@csiro.au (L. Wieczorek).

http://dx.doi.org/10.1016/j.snb.2014.05.094
0925-4005/

classification trees, with each tree being constructed from randomly selected subsets of the data (with replacement), commonly referred to as "bootstrapping". Each data sample is classified by every tree, except for the trees that used that sample in their construction. The output from the classification trees is then averaged to give a prediction, in this case the sensor responses are used to build a "forest" of classification trees and predicted concentrations are output from the trees. The random forests algorithm can handle a large number of different features (descriptors) including redundant and irrelevant features and it does so without prior feature selection and with little or no parameter tuning [37]. Because of the Law of Large Numbers, the random forests algorithm is typically not prone to overfitting and becomes more accurate with more trees [34]. When tested on a wide range of different dimensional data sets, random forests outperformed ANN and SVM in three criteria that assessed the accuracy of the different technique's predictions [38]. However in another study, where there were gaps in the data, random forests underperformed in comparison to multilayered perceptron (MLP) and multivariate adaptive regression spline (MARS) [39]. One study that has examined random forests in relation to sensor array data found that it had a similar classification performance to SVM [7], but random forests is more robust showing little sensitivity to its two hyper parameters [32]. Furthermore, random forests outperformed BPNN and SVM at classifying juices and vinegars that were sampled with an electronic tongue [30]. Random forests analysis of partially selective colorimetric sensor arrays was able to classify whether a patient has cancer with 73.3% sensitivity and 72.4% specificity [33]. This demonstrates that random forests is excellent at classifying samples, but it can also quantify samples. In this paper we use random forests to quantify the concentration of hydrocarbons in water; a challenge necessary to develop a sensor system suitable for environmental monitoring.

Measuring groundwater for volatile organic compounds (VOCs) like benzene, toluene, ethylbenzene and xylene (commonly referred to as BTEX) is of interest at sites where petrochemicals have been stored. The speciation and quantification of BTEX in groundwater can prove challenging because there are potentially hundreds of other hydrocarbon compounds that can simultaneously dissolve into the water from a gasoline or diesel source [40]. The standard method to measure VOCs is by off-site analysis with a gas chromatograph–mass spectrometer (GC–MS) which has excellent detection limits and can accurately quantify the components [41]. Similarly, infrared-attenuated total reflectance (IR-ATR) also shows excellent selectivity and sensitivity [42]. Although these sophisticated instruments are becoming smaller and more portable [43], it is not currently economically feasible to use GC–MS or IR-ATR for constant and real time monitoring of wells. The gold nanoparticle chemiresistor sensor arrays described herein offer a more feasible alternative [44,45]. They have been used to directly detect a variety of hydrocarbons in water [18], seawater [46], and even biological media [47]. Gold nanoparticle films can be used as chemiresistor sensors because their electrical resistivity responds to chemical changes that occur in their immediate environment [48]. Here we examine the performance of chemiresistor sensor arrays to quantify the concentration of multicomponent, synthetic hydrocarbon mixtures in water with the random forests analysis. The concentrations tested are relevant to the EPA maximum contaminant levels in drinking water which are 1 mg/L for toluene, 0.7 mg/L for ethylbenzene and 10 mg/L for *p*-xylene [49].

2. Experimental

Benzene (B), toluene (T), ethylbenzene (E), *p*-xylene (X), and naphthalene (N) 99% reagent grade were used as received from Sigma Aldrich. The combinations of the five analytes that were examined followed a two level, full factorial design. This experimental design was chosen because of its simplicity and ability to completely cover the experimental space of mixture combinations. Every one of the 32 possible combinations of analytes was prepared: one blank (O), 5 unary component mixtures (B, T, E, X and N), 10 binary mixtures (BE, BN, BT, BX, EN, EX, TE, TN, TX and XN), 10 ternary mixtures (BEN, BEX, BTE, BTN, BTX, BXN, EXN, TEN, TEX and TXN), 5 quaternary mixtures (BEXN, BTEN, BTEX, BTXN and TEXN) and one quinary mixture (BTEXN). In addition, a second quinary mixture was prepared that contained all 5 components but at half the nominal concentration of the BTEXN sample, this mixture is referred to as "MID" in the following sections. The first experiment tested mixtures with nominal concentrations of 0 or 10 mg/L of benzene, toluene, ethylbenzene, and *p*-xylene and a nominal concentration of 0 or 5 mg/L naphthalene. For example, the three component "BEN" mixture had a nominal concentration of 10 mg/L benzene, 10 mg/L ethylbenzene, 5 mg/L naphthalene, 0 mg/L toluene and 0 mg/L *p*-xylene. In this experiment the "MID" sample contained benzene, toluene, ethylbenzene and *p*-xylene at a nominal concentration of 5 mg/L and naphthalene at a nominal concentration of 2.5 mg/L. Mixtures were prepared in water from 0.01 g/mL stock solutions of each hydrocarbon in methanol. Possible effects from the presence of methanol were considered, however there were no significant differences in the outcomes from analyses of data that had been corrected for the contribution from methanol in comparison to those that had not. The samples were prepared in a randomized order and then loaded into the fluidic system that delivered them to the array of sensors in eight exposures.

A typical sensor consisted of a gold microelectrode (10 interdigitated fingers, 3000 μm long, 5 μm wide, with a 5 μm gap). On the microelectrode an aqueous solution of gold nanoparticles that are stabilized with 4-(dimethylamino)pyridine are printed and then dried to leave a circular film [51]. The film of nanoparticles is functionalized [44] by incubating the film in an acetonitrile solution with 10 mM of the thiol for 1 h. During incubation the thiol displaces the weakly bonded 4-(dimethylamino)pyridine from the surface of the gold nanoparticles. For the first experiment the array of sensors were functionalized with a variety of thiols: 1-hexanethiol (HEXT), 6-mercapto-1-hexanol (MHOH), 1,8-octanedithiol (OCDT), 2-phenylethanethiol (PET), 2-naphthalenethiol (NAP), (3-mercaptopropyl)triethoxysilane + 1-hexanethiol (MPTES), and trans-4,5-dihydroxy-1,2-dithiane (DOHDT). Each of these thiols except OCDT were used to functionalize two duplicate sensors giving an initial array of 13. The thiols stabilize the nanoparticle and define the partitioning of any chemical into the film from the surrounding environment [52]. Following the incubation, the gold nanoparticle films were rinsed with acetonitrile and water and the baseline resistance measured to confirm the 4-(dimethylamino) pyridine had been displaced.

The second experiment tested all of the same combinations of mixtures as the first experiment, however the nominal concentrations were either 0 or 1 mg/L for each of the five components, including naphthalene, and the "MID" sample contained nominally 0.5 mg/L of each benzene, toluene, ethylbenzene, *p*-xylene, and naphthalene. The mixtures were prepared in a randomized order and delivered in six exposures to the sensor array with the same fluidic system. The sensor array used in the second experiment was functionalized with HEXT ×2, MHOH ×2, OCDT ×2, PET, NAP ×2, 1-octanethiol (OCT), triphenylmethanethiol (TPMT), cyclopentanethiol + 1-hexanethiol (CPT + H), 4-methoxybenzyl mercaptan + 6-mercapto-1-hexanol (MOB + M), and 4-(tert-butyl)benzyl mercaptan + 2-phenylethanethiol (TBBT + PET). Sensors marked with ×2 were present in the array in duplicate otherwise only one of that sensor was present. The MPTES, CPT + H, MOB + M, and TBBT + PET sensors were functionalized with solutions that