Chemosphere 185 (2017) 1063–1071

Contents lists available at ScienceDirect

Chemosphere

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

Assessment of metal ion concentration in water with structured feature selection

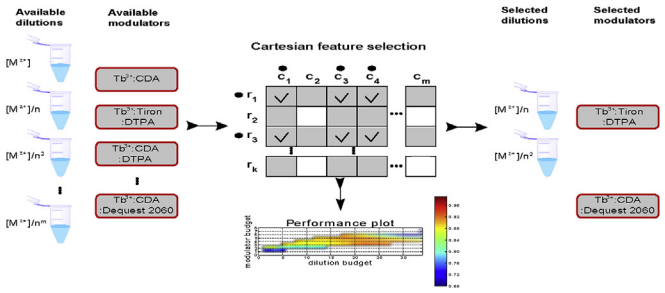
Pekka Naula [a, *], Antti Airola [a], Sari Pihlasalo [b], Ileana Montoya Perez [a], Tapio Salakoski [a], Tapio Pahikkala [a]

[a] Department of Future Technologies, 20014, University of Turku, Finland
[b] Laboratory of Materials Chemistry and Chemical Analysis, Department of Chemistry, 20014, University of Turku, Finland

HIGHLIGHTS

- Contaminated drinking water is a serious problem in developing countries.
- Propose a method for assessing metal ion concentration in drinking water.
- Luminometric label array used for measurements.
- Cost-effective machine learning model for predicting ion concentration.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 10 March 2017
Received in revised form
16 June 2017
Accepted 15 July 2017
Available online 17 July 2017

Handling Editor: Martine Leermakers

Keywords:
Array development
Feature selection
Luminescence
Machine learning
Metal ion quantification
Water analysis

ABSTRACT

We propose a cost-effective system for the determination of metal ion concentration in water, addressing a central issue in water resources management. The system combines novel luminometric label array technology with a machine learning algorithm that selects a minimal number of array reagents (modulators) and liquid sample dilutions, such that enable accurate quantification. The algorithm is able to identify the optimal modulators and sample dilutions leading to cost reductions since less manual labour and resources are needed. Inferring the ion detector involves a unique type of a structured feature selection problem, which we formalize in this paper. We propose a novel Cartesian greedy forward feature selection algorithm for solving the problem. The novel algorithm was evaluated in the concentration assessment of five metal ions and the performance was compared to two known feature selection approaches. The results demonstrate that the proposed system can assist in lowering the costs with minimal loss in accuracy.

1. Introduction

The analysis of household or drinking water and especially the determination of (heavy) metal ion concentration are important due to the safety and customer satisfaction. Generally, the quality of drinking water is high in industrialized countries. However, according to UNICEF, 1.1 billion people mainly in part of the African

* Corresponding author.
E-mail addresses: pekka.naula@utu.fi (P. Naula), antti.airola@utu.fi (A. Airola), saanpi@utu.fi (S. Pihlasalo), iimope@utu.fi (I. Montoya Perez), tapio.salakoski@utu.fi (T. Salakoski), tapio.pahikkala@utu.fi (T. Pahikkala).

http://dx.doi.org/10.1016/j.chemosphere.2017.07.079
0045-6535/

and Asian countries lack access to improved drinking water sources. Many methods have been proposed for the determination of metal ion concentrations from water. However, methods such as atomic absorption spectrometry (PerkinElmer, 2011), inductively coupled plasma atomic emission spectrometry (PerkinElmer, 2011), X-ray fluorescence spectrometry (Kot et al., 2000; Panayappan et al., 1978), and polarography (Jakumnee et al., 2002; Babaei et al., 2007) methods, require transfer to laboratory, costly instruments, use of toxic mercury, preconcentration steps to achieve high sensitivity, and/or special expertise. With ion selective electrodes (Bakker and Pretsch, 2008), the drawback is a specific electrode that is needed for each metal ion. Moreover, colorimetric, spectrophotometric, and fluorometric methods have been developed for the detection of metal ions. They suffer from the interfering metal ions and need for different assay protocols, incubation times, and reagents for each metal ion (Pihlasalo et al., 2016). Fluoroionophores utilized in fluorometric methods require also long and laborious synthesis. For a more detailed review of relevant methods, see our previous article (Pihlasalo et al., 2016), as well as the review article of Pesavento et al. (2009).

We have developed a novel label array for the determination of metal ion concentrations in liquid samples (Pihlasalo et al., 2016), which would allow the determination of several metal ions by utilizing different mixtures of array reagents (modulators). In this article, we develop a structured feature selection method applicable for selecting the optimal modulators and dilutions for such label arrays. These arrays are not applicable only for quantification of metal ions. Instead, they are suitable for various identification and mixture analysis tasks, development and fine-tuning of products for customers, analysis of authenticity, detection of product adulteration (Härmä et al., 2015), and quality control of liquid and liquidizable samples.

Both the different modulator types and dilution ratios used in the array come with a cost. Each modulator corresponds to a set of labels and additional modulator reagents added to a well of a microtiter plate, and thus it pays to minimize the number of reagents required for the experiment. Similarly, each dilution increases the manual work required. Thus it would be beneficial, that for any given ion detection task we could automatically find a minimal subset of the possible modulators and dilutions, such that allow assessing the concentration accurately.

In a feature selection process, irrelevant and redundant features are removed from the set of all possible features. Such selection can be performed for a number of reasons, including the prevention of overfitting, ability to obtain simple models interpretable by human experts, and in order to reduce the cost of measuring feature values. Our focus in this work is on the last of these three criteria. Typically, the cost sensitive feature selection problems are considered as feature selection with a budget, that is, the number of features the model can depend on is restricted and the aim is to maximize the prediction performance under this constraint (see Xu et al. (2012); Naula et al. (2014)). A large variety of feature selection methods have been proposed in the literature (see e.g. Guyon and Elisseeff (2003) for an overview), including the usage of statistical prefilters, L1-regularization, and wrapper based search methods that select features based on prediction error typically estimated using cross-validation. However, standard feature selection methods are not able to properly model settings, where each feature is formed by combining elements from two distinct sets, such as in the case of dilution-modulator combinations in the application considered here.

In order to develop a method that is able to select the optimal modulators and dilutions, we implement a Cartesian feature selection method. For conventional feature selection problems, greedy methods have been recently shown in a comprehensive experimental comparison (Pahikkala et al., 2012b; Naula et al., 2014) to have state-of-the art performance in settings where the number of features needs to be restricted to be as low as possible. Further, greedy methods are known to be applicable to enforcing more complex structured sparsity patterns on learned models (see e.g. Huang et al. (2011)).

Selection of optimal sensors for classification or regression is a problem that has been studied in a large variety of different application domains (see e.g. Alstrøm et al. (2011)), with most works considering the standard feature selection problem where no special structure is present. Recently, Nowotny et al. (2013) considered a feature selection setting with Cartesian structure, where optimal combination of metal oxide sensors and sampling times was selected for classification of chemicals using a linear model. However, their work did not formalize the Cartesian feature selection problem, or propose an efficient algorithm for minimizing the Cartesian feature costs.

The main contributions of this article are to

- formally define Cartesian feature selection problem
- test three feature selection algorithms for solving it, the proposed Cartesian greedy method, as well as two simpler adaptations of existing methods
- show that Cartesian feature selection allows accurate prediction of metal ion concentration from water with low number of modulators and dilutions

2. Methods and materials

2.1. Cartesian feature selection

Given the luminescence signals as an output by the label array, the aim is to determine the metal ion concentration in water. Each modulator is applied to a sample of water possibly diluted with a given ratio and the luminescence signals are monitored. The feature representation for the data is thus formed as follows. First, we have available an array of modulators. Then, all of these modulators are applied to different dilutions of the water sample being analyzed

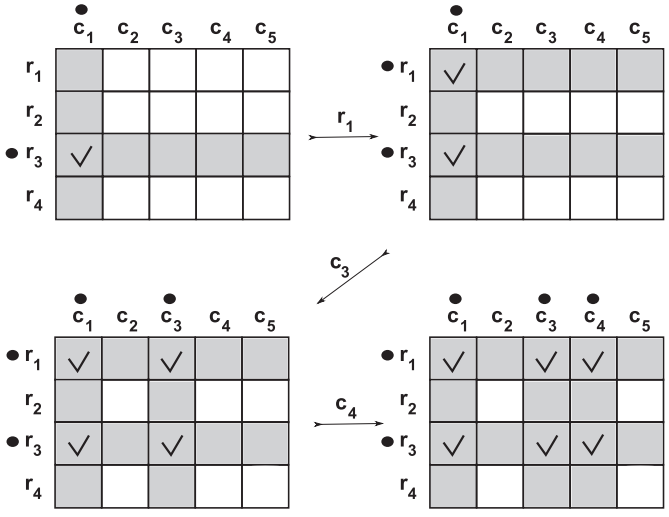


Fig. 1. Top left: The algorithm is initialized by selecting first the feature providing the lowest LOOCV error, in this case (r_3,c_1). Top right: one row index becomes selected and now altogether two features have been selected. Bottom left: one column is selected, now the set of selected features consists of four features e.g. the size of the set is increased by 2 for the price of a single index addition. Bottom right: one additional column is selected thus giving two more features by the cost of one index addition.