# 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*

**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.

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

## 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 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 se-

*Corresponding author. Tel.: +358 405022708; fax: +358 2 2410154

*Email addresses:* pekka.naula@utu.fi (Pekka Naula), antti.airola@utu.fi (Antti Airola), saanpi@utu.fi (Sari Pihlasalo), iimope@utu.fi (Ileana Montoya Perez), tapio.salakoski@utu.fi (Tapio Salakoski), tapio.pahikkala@utu.fi (Tapio Pahikkala)

lection 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 pre-filters, 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 (including also the undiluted sample).

The set of features describing the data consists of the Cartesian product of the set of modulators

and the set of dilution ratios. The resulting feature representation, here referred to as the Cartesian features, is visualized in Fig. 1. Here, the rows of the feature grid can be mapped to the modulators, the columns to the dilutions, and the elements to the feature values measured by a given modulator when applied to a given dilution. Given $|\mathcal{R}|$ modulators and $|\mathcal{C}|$ dilutions, the overall number of features describing a water sample is thus $|\mathcal{R}||\mathcal{C}|$. As an example, let $\mathcal{R} = \{r_1, r_2, r_3, r_4\}$ and $\mathcal{C} = \{c_1, c_2, c_3, c_4, c_5\}$. Their Cartesian product consists of $4 \times 5 = 20$ features, a single feature denoted here as $(r_i, c_j)$.

Inferring the metal ion concentration from these signals can be modeled as a standard regression problem. Given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$ of water samples, where $\mathbf{x}_i \in \mathbb{R}^{|\mathcal{R}||\mathcal{C}|}$ is the feature vector encoding the modulator-dilution combinations for a given water sample, and $y_i \in \mathbb{R}$ the metal ion concentration, we aim to find a predictor $f(\mathbf{x}) \approx y$, that can accurately predict metal concentration for a water sample not present in the training set.

We assume that each index in both sets is associated with a cost. For the sake of simplicity in this work we assume that each index has the same unit cost. Further, we assume a fixed budget on the total cost allowed. This setting gives rise to the Cartesian feature selection problem, whose search space consists of tuples $(P, Q)$ such that $P \subseteq \mathcal{R}$, $Q \subseteq \mathcal{C}$ and $|P| + |Q| \leq k$ that satisfy this budget $k$. The problem is to find such tuple, whose corresponding feature set determined by $P \times Q \subseteq \mathcal{R} \times \mathcal{C}$ maximizes the prediction performance.

To solve this problem, we propose a search algorithm based on a greedy forward selection heuristic (Algorithm 1). We follow the so-called wrapper approach (see e.g. Kohavi and John (1997)) to feature selection, where the objective function $J$ used for selection is the cross-validation (CV) error obtained with a learning algorithm trained using a given set of feature indices. The first feature selected is the one that provides the lowest CV error (line 3). On line 7 the algorithm evaluates the CV error for each index $r$ in $\mathcal{R}$ not yet added into $P$ by calling the objective function $J$ with the feature set $(P \cup \{r\}) \times Q$. The same is done on line 8 for $\mathcal{C}$. The index providing the lowest CV error is selected on each round until $k$ indices have been selected.

Fig. 1 demonstrates the progress of the algorithm. The matrix contains the overall set of available features and its rows and columns are indexed by $\mathcal{R}$ and $\mathcal{C}$. The set of selected feature indices after the three steps is $\{r_1, r_3\} \times \{c_1, c_3, c_4\}$. Note that the two last steps in Fig. 1 each increase the size of the set of selected features by two while the budget is only increased by one. This effect is further emphasized, for example, when the size of $|P|$ is large, since each index addition to $Q$ increases the number of selected features by $|P|$.

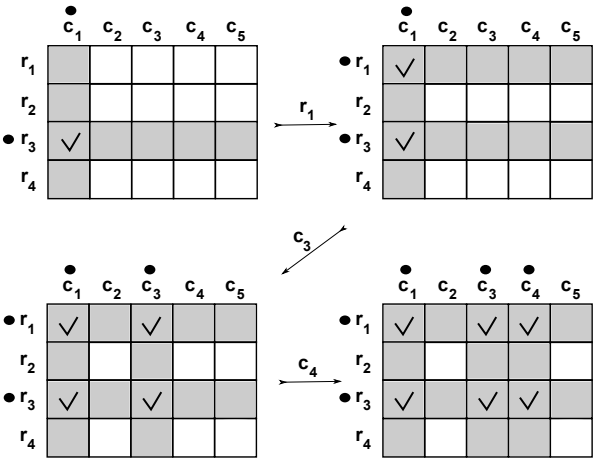


Figure 1: Top left: The algorithm is initialized by selecting first the feature providing the lowest LOOCV error, in this case ($r_3$,$c_1$). Top right: one row index becomes selected and now altogether two features have been selected. Bottom left: one column is selected, now the set of selected features 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.

In our experiments, we use two well-known learning algorithms, the k-nearest neighbors regression (kNN) (Cover, 1968) and ridge regression. When making a prediction for new data point, kNN finds the $k$ data points nearest to it (Euclidean distance), and predicts the mean of their outputs. Ridge regression infers a linear prediction function $\mathbf{w}$ that minimizes the following function:

$$\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|^2,$$

where $\mathbf{X}$ is the data matrix, $\mathbf{y}$ is a vector containing the corresponding outputs and $\lambda$ is the ridge penalty parameter. Note also that the data matrix is always assumed to include a single constant valued feature that simulates the so-called intercept term in linear models and that is not involved in the feature selection process.

An advantage of both of the methods is that cross-validation error can be computed efficiently. For ridge regression this can be accomplished by us-

**Algorithm 1** Cartesian greedy forward selection

$\mathcal{R} \neq \emptyset$ ▷ Index set 1
$\mathcal{C} \neq \emptyset$ ▷ Index set 2
$(a, b) \leftarrow \text{argmin}_{(r,c)\in\mathcal{R}\times\mathcal{C}} J((r, c))$ ▷ Initialize first index pair
$P \leftarrow \{a\}$ ▷ Set of selected indices from $\mathcal{R}$
$Q \leftarrow \{b\}$ ▷ Set of selected indices from $\mathcal{C}$
**while** $|P| + |Q| < k$ **do**
    $a \leftarrow \text{argmin}_{r\in\mathcal{R}\setminus P} J((P \cup \{r\}) \times Q)$
    $b \leftarrow \text{argmin}_{c\in\mathcal{C}\setminus Q} J(P \times (Q \cup \{c\}))$
    **if** $J((P \cup \{a\}) \times Q) < J(P \times (Q \cup \{b\}))$ **then**
        $P \leftarrow P \cup \{a\}$
    **else**
        $Q \leftarrow Q \cup \{b\}$
**Return** $P$, $Q$

ing the classical Woodbury matrix inversion identity (Wahba, 1990), whereas for kNN this can be implemented using for example the k-d tree data structure to speed up neighbor searches (Bentley, 1975).

### *2.2. Data*

Table 1 summarizes the characteristics of the data sets. There are in total 11 datasets each containing one of five different metal ions ($Cd^{2+}$, $Pb^{2+}$, $Cu^{2+}$, $Fe^{2+}$, and $Ni^{2+}$). For each concentration four replicas are measured. Each data point corresponds to one such replica, thus the number of data points for each data set is four times the number of concentrations. For $Cd^{2+}$, $Pb^{2+}$, $Cu^{2+}$, and $Fe^{2+}$ there are two time points, and for $Ni^{2+}$ three, corresponding to time when the modulator signal has been measured (see Table 1). The set of available modulators for the different datasets depends on the type of the metal.

The modulator solutions were added as two separate 3.0 µL droplets in MilliQ water and dried to the black polystyrene 96-well microtiter plates. Modulator solutions contained 5.0 µM $TbCl_3$ ($TbCl_3 \cdot 6\,H_2O$ from Sigma-Aldrich (St. Louis, MO)) in the first droplet and different ligands (chelidamic acid (CDA) hydrate, diethylenetriaminepentaacetic acid (DTPA), and diethylenetriaminepentakis (methylphosphonic acid) (Dequest 2060) from Sigma-Aldrich (St. Louis, MO) and 4,5-dihydroxy-1,3-benzenedisulfonic acid (Tiron) disodium salt monohydrate from Acros Organics (Geel, Belgium)) in the second droplet as described in Table 1. Furthermore, additional modulators (HCl from Sigma-Aldrich (St. Louis, MO) and citric acid monohydrate from Merck KGaA (Darmstadt, Germany)) were added to the sample in modulators 1-3. The metal ions spiked as chloride or sulfate salts ($CuSO_4 \cdot 5H_2O$, $FeSO_4 \cdot 7H_2O$, and $NiCl_2 \cdot 6H_2O$ from Sigma-Aldrich (St. Louis, MO), $CdCl_2 \cdot H_2O$ from Merck KGaA (Darmstadt, Germany), and $PbCl_2$ from Acros Organics (Geel, Belgium)) to household water (from an apartment house in the Turku city center) were pipetted in 100 µL to the modulator plates. Terbium luminescence emission intensities were measured with four replicates in a 400 µs window after a 400 µs delay time with 320 nm excitation and 545 nm emission wavelengths using a Labrox plate reader (Labrox, Turku, Finland) at different incubation times after the addition of the sample to the wells. The metal ion concentrations in Turku household water are as follows: Cd $< 0.01$µg/L, Cu 0.55µg/L, Fe $< 5$µg/L, Pb $< 0.05$µg/L and Ni $< 0.3$µg/L (Turku Region Water Ltd., 2016).

## 3. Results and discussion

In the computational experiments, we test on a number of real-world data sets feature selection approaches for selecting jointly modulators and dilutions necessary for accurate determination of metal ion concentration in water. We compare the Cartesian greedy forward selection algorithm to two simpler alternative approaches; one based on standard greedy forward selection (Pahikkala et al., 2012b; Naula et al., 2014), and one on the Lasso approach (Tibshirani, 1996). None of these methods have previously been applied to Cartesian feature selection problems. Further, we analyze how the accu-

Table 1: Characteristics of the data sets (top), incubation times for the assessment of metal ion concentration and selected modulators for a medium budget (middle), and modulator solutions (bottom).

| Metal ion | $Cd^{2+}$ | $Pb^{2+}$ | $Cu^{2+}$ | $Fe^{2+}$ | $Ni^{2+}$ |
|---|---|---|---|---|---|
| Time points | 2 | 2 | 2 | 2 | 3 |
| Concentrations | 26 | 26 | 27 | 27 | 36 |
| Data points | 104 | 104 | 108 | 108 | 144 |
| Concentration range (µg/L) | 20–5100 | 20–5100 | 9.3–3000 | 9.3–3000 | 1.2–3000 |
| Modulators | 4,6,7 | 4,6,7 | 1–5 | 1–5 | 4–11 |
| Dilutions | 24 | 24 | 25 | 25 | 34 |

| Dataset | Incubation times for modulators | Selected modulators |
|---|---|---|
| $Cd^{2+}$, timepoint 1 | 4: 40 min, 6: 1 h, 7: 40 min | 4, 7 |
| $Cd^{2+}$, timepoint 2 | 4: 3 h, 6: 4 h, 7: 3 h | 4, 7 |
| $Pb^{2+}$, timepoint 1 | 4: 40 min, 6: 1 h, 7: 40 min | 7, 4 |
| $Pb^{2+}$, timepoint 2 | 4: 3 h, 6: 4 h, 7: 3 h | 7, 4 |
| $Cu^{2+}$, timepoint 1 | 1: 1 h, 2: 2 h, 3: 1 h, 4: 1 h, 5: 45 min | 3, 4 |
| $Cu^{2+}$, timepoint 2 | 1-4: 3 h, 5: 1 h | 3, 4, 5, 2 |
| $Fe^{2+}$, timepoint 1 | 1: 1 h, 2: 2 h, 3: 1 h, 4: 1 h, 5: 45 min | 2, 3 |
| $Fe^{2+}$, timepoint 2 | 1-4: 3 h, 5: 1 h | 2 |
| $Ni^{2+}$, timepoint 1 | 4-11: 20 min | 5, 9, 8 |
| $Ni^{2+}$, timepoint 2 | 4-11: 1 h | 5, 6, 8, 9, 4, 11 |
| $Ni^{2+}$, timepoint 3 | 4-11: 2 h | 5, 6, 8, 4, 9 |

| Modulator | Ligand 1 | Ligand 2 | Additional Modulator |
|---|---|---|---|
| 1 | 20 µM CDA | 10 µM Dequest 2060 | 27 µM citric acid |
| 2 | 50 µM Tiron | | 27 µM citric acid |
| 3 | 20 µM CDA | 20 µM DTPA | 300 µM HCl |
| 4 | 20 µM CDA | 20 µM DTPA | |
| 5 | 20 µM CDA | 10 µM Dequest 2060 | |
| 6 | 500 µM CDA | 2 µM Dequest 2060 | |
| 7 | 50 µM Tiron | | |
| 8 | 50 µM Tiron | 20 µM DTPA | |
| 9 | 50 µM CDA | 50 µM DTPA | |
| 10 | 50 µM CDA | | |
| 11 | 50 µM Tiron | 20 µM Dequest 2060 | |

racy of the learned model changes as a function of both the number of selected dilutions and modulators, and present the optimal modulators selected in each experiment.

### *3.1. Feature representation*

Metal ions are detected from water with a novel nonspecific label array (Pihlasalo et al., 2016) that comprises a set of chemical modulators. The dilutions with different dilution ratios were prepared from the water sample containing the spiked metal ions. The original and diluted samples were measured with the set of modulators. As an example, let us consider the concentrations of $Cd^{2+}$ that were measured with three modulators. The responses of the three modulators as a function of $Cd^{2+}$ concentration are illustrated in Fig. 2. We observe that the signal values do not change linearly with respect to the concentration, and hence the determination from the original water sample only is challenging. In contrast, the determination is considerably more accurate if the signal can be measured from different ranges of the modulator response curves, especially from the dynamic range of each modulator. This is achieved via dilution of the original sample with different ratios. For example,

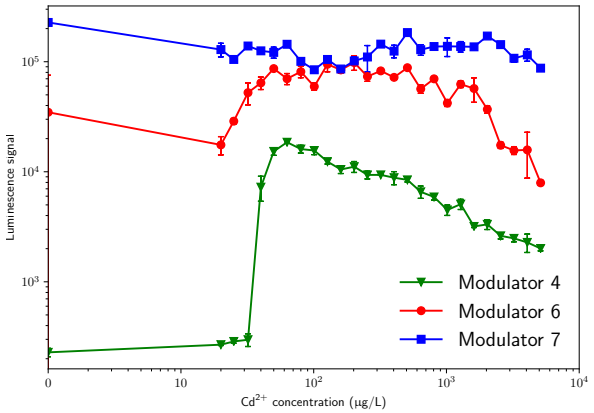


Figure 2: Signal measured with three modulators at varying concentrations of $Cd^{2+}$. The markers and bars represent, respectively, the means and standard deviations of the four replicas.

if we have two water samples with $Cd^{2+}$ concentrations 40 µg/L and 640 µg/L (the 5th and 17th measurement points in Fig. 2), the corresponding signal values measured with modulator 4 are, respectively, 7245.5 and 6570.75, making them hardly distinguishable, because the signal value first in-

creases and then starts to decrease again as a function of concentration. However, diluting the samples with ratio 0.5 provides us additional modulator signal values 287.0 and 15598.0 corresponding to the concentrations 20 µg/L and 320 µg/L (the 2nd and 14th measurement points in Fig. 2) for the two samples, which together with the original measurements clearly differentiates the two samples. Thus, each water sample is transformed to a feature vector consisting of the modulator signal values with a set of different dilution ratios. The transformation is nonlinear in the sense that it is determined by the underlying modulator signal curves, and hence it reflects the nonlinear nature of the signals.

### *3.2. Setup*

We test the proposed Cartesian greedy feature selection method (Algorithm 1). The Cartesian method directly optimizes the considered cost by selecting at each step a new dilution or modulator, and then uses all the features corresponding to the Cartesian product of the selected dilutions and modulators.

Additionally, we compare the proposed method to two simpler adaptations of existing methods, that do not directly optimize the cost. These two approaches select features directly from the pool of dilution-modulator pairs, without taking advantage of the Cartesian structure. Considering Figure 1, the Cartesian method selects row- or column indices of the matrix, while the other methods select directly elements of the matrix. The Cartesian method has the advantage that it can more directly minimize the combined cost of dilutions and modulators, whereas the other two methods have the advantage that they can be implemented using existing feature selection methods and software implementations.

The first alternative method is greedy forward selection (see e.g. (Pahikkala et al., 2010)). The second alternative feature selection method applied is the $l_1$-regularized lasso method (Tibshirani, 1996). Following the procedure described by Friedman et al. (2009), the final models for lasso-selected features are subsequently re-fitted with $l_2$-ridge regression method. This was done because we noticed in preliminary experiments, that using directly the $l_1$-regularized lasso model gave very poor prediction performance, because the heavy regularization required for obtaining small feature sets also causes a strong bias on the inferred model (Zhang, 2011).

The Cartesian and greedy feature selection was implemented both using kNN and ridge regression as the learning algorithm. The kNN methods were implemented using the KDTree module in the scikit-learn library (Pedregosa et al., 2011). The ridge regression methods were implemented using the RLS and greedyRLS modules in the RLScore machine learning library (Pahikkala and Airola, 2016). Implementation of the LARS algorithm (Efron et al., 1996) in scikit-learn was used for Lasso feature selection.

The number of neighbors $k$ for kNN is set to 8, based on our prior knowledge that the training set contains 4 replicas for each measured concentration. Then, given a new concentration value $y$ to be predicted, the average value of a set of eight neighbors containing the four largest concentration values in the training set that are smaller than $y$ and the four smallest values that are larger than $y$ is likely to be close to $y$. The heuristic used for feature selection with kNN was leave-concentration-out CV, where all the 4 replicas corresponding to a single concentration were left out on each round.

Ridge regression contains the regularization parameter $\lambda$ that adjusts the level of shrinking of the model coefficients. We select the parameter from grid $[2^{-20}, 2^{-18}, \ldots, 2^{20}]$ in an inner leave-one-out CV process on the four training folds and the greedy methods use the same leave-one-out estimates as the heuristic for feature selection. That is, the test folds for all rounds of CV are only used for prediction performance evaluation.

As a summary, each data set contains 104-144 data points with $\mathit{modulators} \times \mathit{dilutions}$ features each. The aim is to predict the concentration levels of different metal ions using at most a given number of modulators and dilutions in order to reduce prediction costs. In order to reduce variance in the results we repeated the following procedure 100 times and calculated the average of the performances. We performed five-fold CV, where each group of four replicas was assigned to the same fold. This assignment was done in order to simulate the typical prediction time situation in which the concentration value to be predicted is never exactly the same as any of the ones used for training the predictor. We stress that in order to avoid the selection bias in prediction performance estimates, the CV used for feature and regularization parameter selection was always performed on the four training folds of the five-fold CV, resulting in a nested CV setting (Varma and Simon, 2006).

To compare the prediction performances of the different regressors, we use the concordance index (C-index) (Gönen and Heller, 2005):

$$\frac{1}{|\{(i,j) \mid y_i > y_j\}|} \sum_{y_i > y_j} H(f(x_i) - f(x_j))$$

$$\text{with } H(d) = \begin{cases} 1 & \text{if } d > 0 \\ 0.5 & \text{if } d = 0 \\ 0 & \text{if } d < 0 \end{cases}$$

This measure was chosen for the following reasons. Firstly, as a rank correlation measure, it is far more stable with small data sets containing outliers than other regression measures, and enabled us to observe the big picture about the performance differences with respect to the different budget values. Secondly, in contrast to the other well-known rank correlation measures such as the Spearman correlation, C-index does not account the pairwise prediction differences between the replicas, making it especially suitable for our data due to the four replicas per each concentration value. The C-index extends the area under ROC curve (AUC) to ordinal and real-valued scales. It measures the ability of the learned model to rank water samples from lowest to highest ion concentration. It obtains value 0.5 for purely random predictions, and 1.0 for perfectly correlated predictions.

### *3.3. Results for data analysis*

In Fig. 3 we show how the Cartesian greedy feature selection method, standard greedy forward selection and lasso compare to each other. The results are presented as curves plotting combined dilution and modulator budget against C-index. Fig. 3 plots the average performance of cross-validation results for 100 repetitions over several budgets. Clearly, Cartesian and greedy outperform lasso with respect to all budgets in visual analysis (see Fig. 3). The Cartesian selection approach appears to also perform slightly better than the standard greedy approach on most data sets. The kNN methods clearly outperform the linear methods.

We tested whether the methods differ statistically significantly for two fixed budget sizes (low=5, medium=15) using the Wilcoxon signed-rank test computed over the 11 data sets ($p < 0.05$). For both budgets, Cartesian kNN outperformed all the other methods significantly ($p = 0.0033$ for all comparisons). Greedy kNN outperformed the ridge and lasso methods ($p = 0.0044$ for greedy ridge and low budget, $p = 0.0033$ for other comparisons). Cartesian ridge significantly outperformed greedy ridge for small budget size ($p = 0.016$), but not medium ($p = 0.53$). Both Cartesian and greedy ridge always outperformed lasso ($p = 0.0033$). To conclude, the nonlinear kNN methods work significantly better than the linear methods, and in most cases Cartesian feature selection outperformed greedy selection. The lasso approach always performed worst of the methods.

To inspect the prediction performance of the best performing method (Cartesian kNN) in more detail, we compute the root mean-squared error (RMSE) separately for each concentration value. The values are computed against the $\log(y + 1)$ transformed concentration values to make the errors more comparable with each other, so as to indicate that larger prediction errors are tolerated for larger true concentration values. These results are illustrated in Figure 4. We observe the relatively large errors on the zero concentrations but this not surprising due to the nature of our concentration-level cross-validation design, as the predictions are always performed with a model trained without the concentration values to be predicted. When applying the model in real world settings, also zero-concentration samples would be present in the training data, which would likely eliminate this problem. The errors also increase with large budgets, due to the vulnerability of kNN to the curse of dimensionality (Radovanović et al., 2010).

Heat maps that provide an overview of the Cartesian kNN C-index results for different modulator and dilution budgets are presented in Fig. 5. The heat map visualizes the predictive accuracy, the models can achieve subject to different budget constraints both on the number of dilutions and modulators. For each round of 100 times repeated 5-fold cross-validation, we performed Cartesian greedy forward selection, until all the modulators and dilutions had been selected. There are thus altogether 500 separate feature selection runs for each data set. Each run corresponds to a path in the heat map. The path starts at the lower left corner, where the initial feature has been selected. After the first step, the path traverses towards the upper right corner, at each step selecting either one dilution or modulator, until all have been selected. Naturally, the paths differ for each of the 500 runs, though for many budget combinations they overlap. In the heat maps, the color denotes the C-index averaged over the paths crossing the given budget combina-

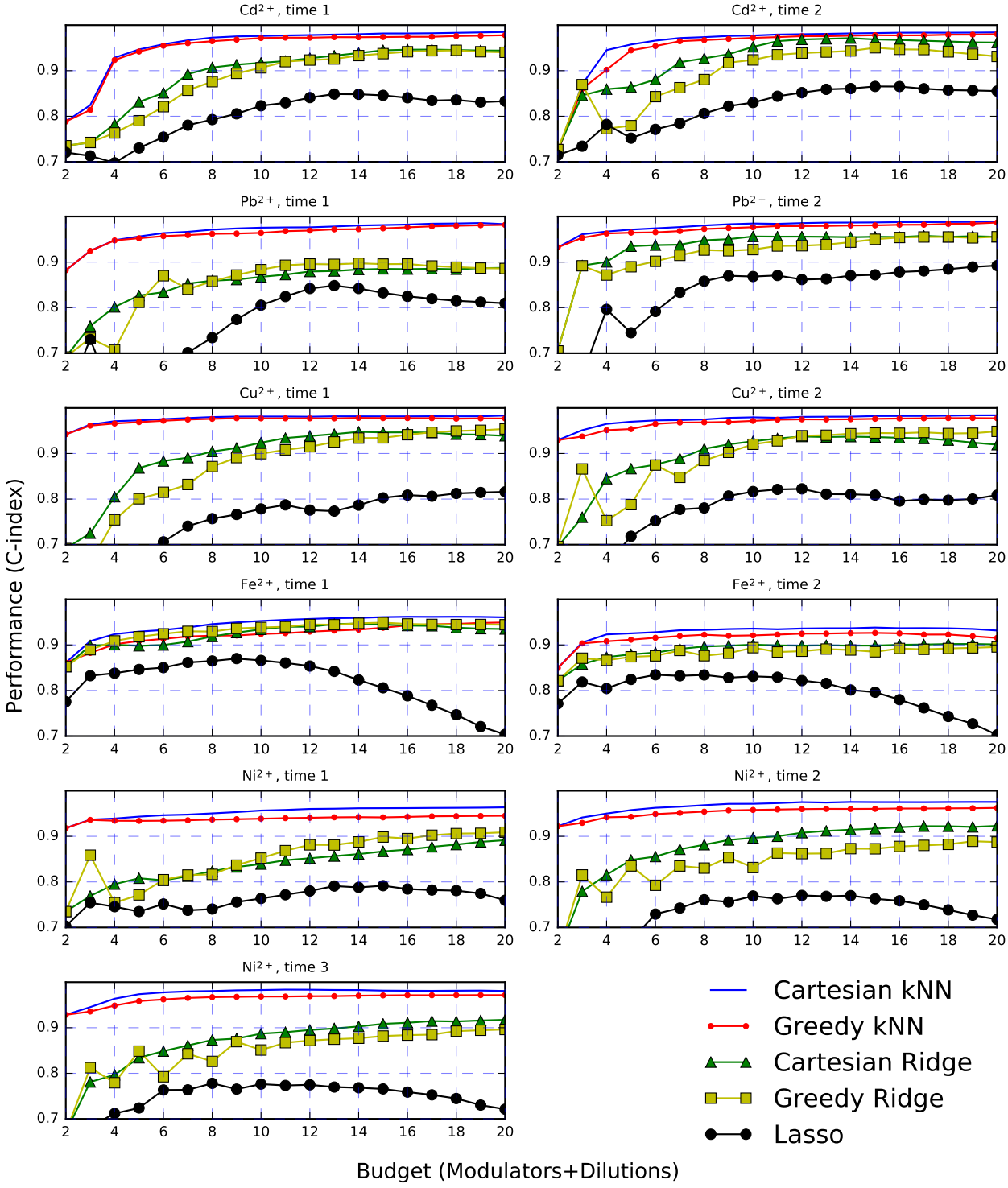


Figure 3: Performance curves of Cartesian, Greedy and Lasso methods on all data sets.

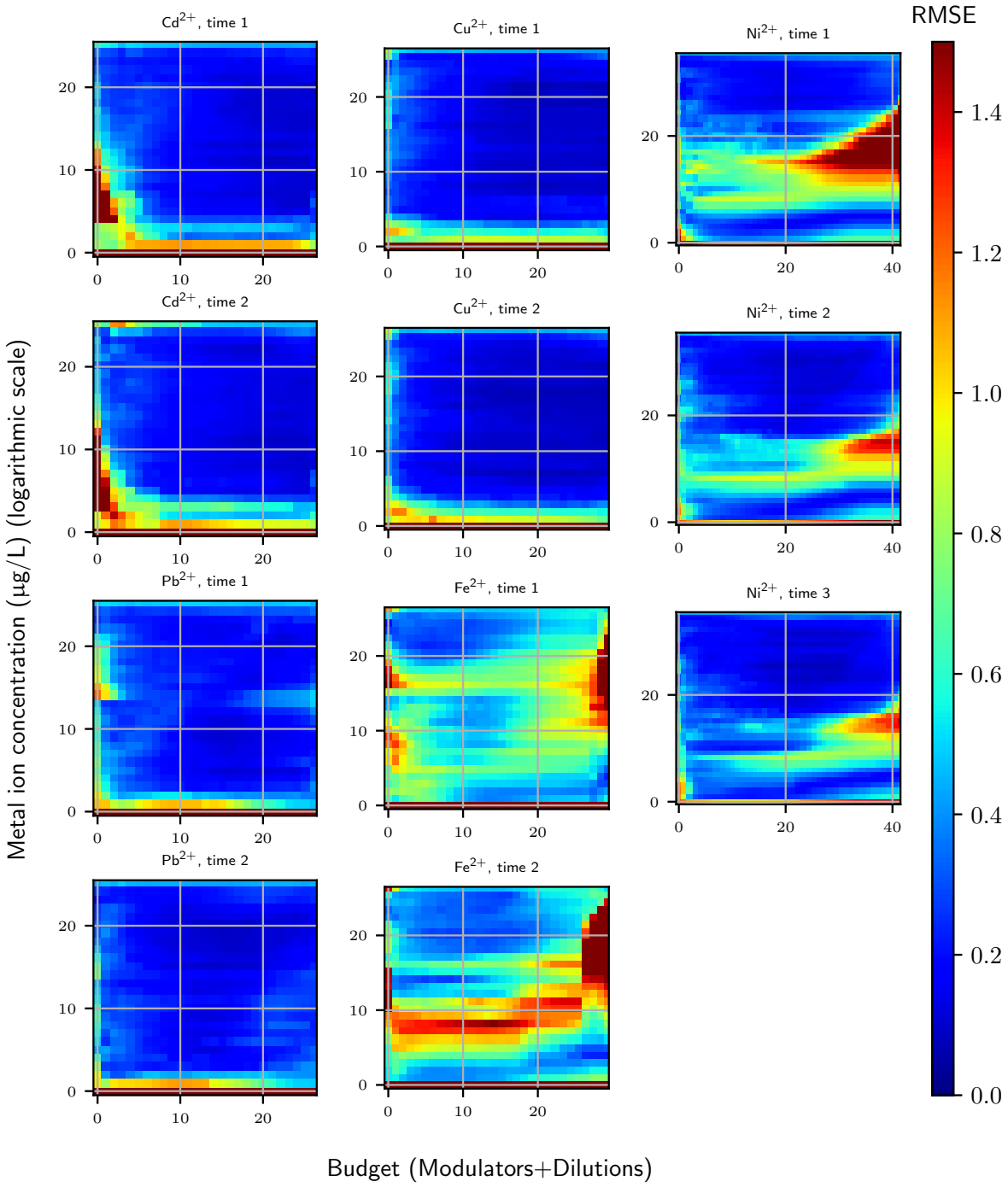


Figure 4: Heat maps illustrating RMSE for each concentration level with different total budget values for the 11 data sets.

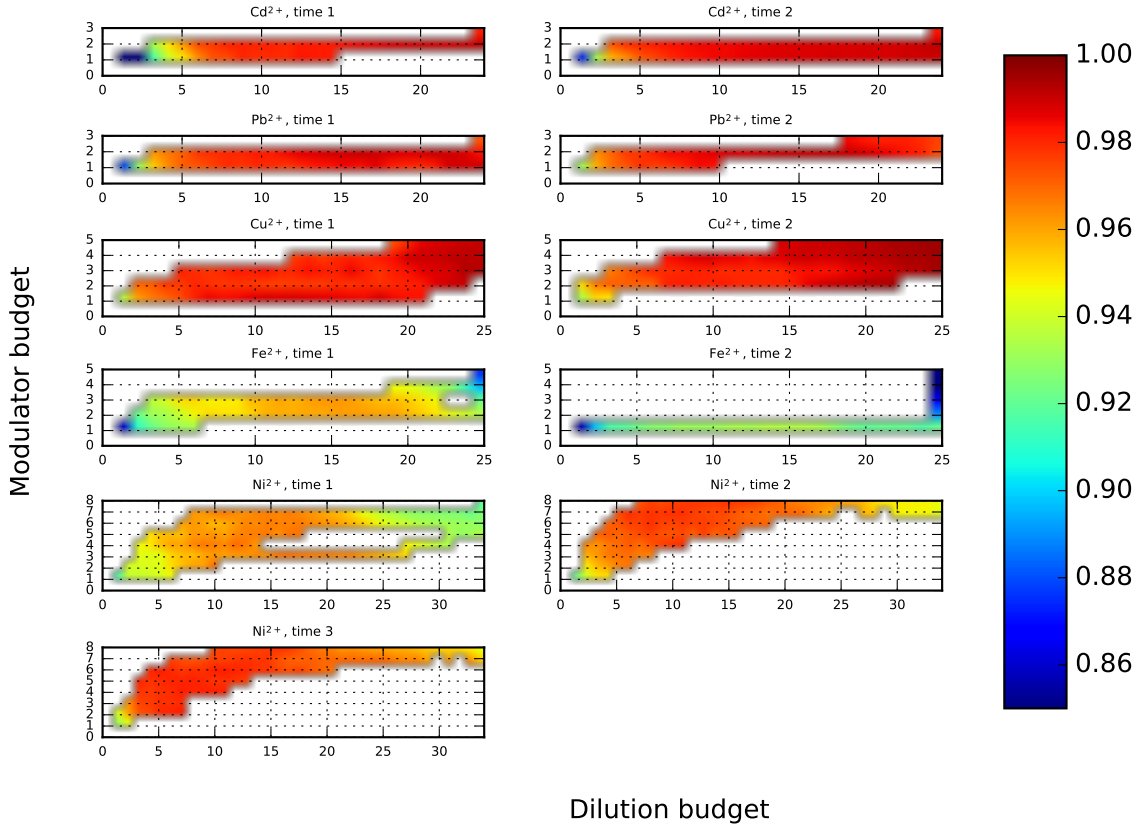


Figure 5: Heat maps illustrating the C-index with different amounts of modulators and dilutions for the 11 data sets.

tion. Areas crossed by less than 50 paths are left white in the heat map in order to filter out noise.

As expected, for the smallest budgets the C-index values are low. Increasing the budgets also results at first in a sharp increase in the predictive performance. However, while this is very data set specific, it can be observed that usually the highest performance is already reached with the middle budget values. The low performance that appears for the $Ni^{2+}$ data sets in the highest budgets demonstrates the benefits that feature selection can sometimes have on the accuracy of the learned predictor. The main findings are however, that the accuracy is highly dependent on the size of the allocated budget, and that high accuracy metal ion detectors can be obtained with moderate costs compared to the case in which no structured feature selection is performed.

In Table 1 we present the modulators selected for a budget of size 10 with Cartesian kNN, the predictive performance does not much improve for any of the data sets for larger budget sizes. The modulators are presented in the order they have been selected, first the most important, second the most important given the first has been selected etc. First, we can see that none of the data sets require all the modulators for accurate prediction, for example for $Ni^{2+}$ timepoint 1 and only three of the eight modulators is selected. The selected modulators are exactly the same for both time points for $Cd^{2+}$, $Pb^{2+}$. For the rest of the datasets the number of selected modulators may differ across time points, but the included modulators, and the order in which they are selected are in most cases the same.

The data set in this study is small avoiding comprehensive conclusions for the selection of optimal modulators. However, the results obtained for

$Cu^{2+}$ suggest that $Cu^{2+}$ can be quantified with the highest accuracy at low pH by utilizing modulator 3 containing DTPA as the nonantenna ligand. Thus, the accuracy is higher, when $Cu^{2+}$ is coordinated to DTPA compared to the coordination to Dequest 2060. This might be related to the lower affinity of DTPA to $Cu^{2+}$ compared to Dequest 2060. The optimal modulators for $Fe^{2+}$ are also well understandable, as $Fe^{2+}$ competes with $Tb^{3+}$ in binding to nonantenna ligands at higher concentrations compared to the binding competition to antenna ligand Tiron. Therefore, the dynamic range is broader with modulator 2 than with modulators 1 and 3-5.

## 4. Conclusions and future work

In this study, we introduced a simple method for making predictions of the quality of drinking water, using a minimal number of modulators and dilutions. The approach is based on a novel Cartesian feature selection algorithm. Our experimental results on 11 data sets show that the algorithm is able to infer accurate ion detectors with a negligible cost compared to the ones obtained using the whole feature set. The advantages of Cartesian selection over other approaches were shown both for kNN based non-linear models, as well as for ridge regression based linear models.

When applying a trained metal ion detection system in new environments, it might be necessary to adapt the system by re-training it on samples gathered from the new environment. In such a situation, the costs of gathering data could be reduced by using only those modulators and dilutions that were chosen on the data used for training the initial system. Further, by applying online learning methods (Pahikkala et al., 2012a), one could start by using a system trained on data from different environment, and then adapt it over time as more data was gathered.

In the future, we intend to extend the experimental analysis with samples from various application fields and larger data sets of mixture samples including multiple metal ions simultaneously. Further, the cost optimization could be improved with more sophisticated search algorithms based on, for example, evolutionary algorithms. One could also design a regularizer structured analogously to the Cartesian cost profile considered in this paper. This could be done, for example, with a variation of the $l_{1,2}$ norm often used in multi-task or grouped feature selection. Namely, the Cartesian sparsity structure can be encouraged via minimizing the sum of both the row-wise and column-wise 2-norms of the modulator-dilution matrix. However, due to the poor results of the ordinary lasso compared to the greedy approaches, we decided to leave this to a future studies. Finally, further application possibilities of the Cartesian feature selection will be explored.


## Acknowledgment

This work was supported by the funding from the Academy of Finland (Grants 289903 and 258617). We would like to thank CSC, the Finnish IT center for science, for providing us with extensive computational resources.


## References


Alstrøm, T.S., Larsen, J., Kostesha, N.V., Jakobsen, M.H., Boisen, A.: Data representation and feature selection for colorimetric sensor arrays used as explosives detectors. In: Tan, T., Katagiri, S., Tao, J., Nakamura, A., Larsen, J. (eds.) 2011 IEEE International Workshop on Machine Learning for Signal Processing. pp. 1–6. IEEE (2011)

Babaei, A., Babazadeh, M., Shams, E.: Simultaneous determination of iron, copper, and cadmium by adsorptive stripping voltammetry in the presence of thymolphthalexone. Electroanalysis 19(9), 978–985 (2007)

Bakker, E., Pretsch, E.: Nanoscale potentiometry. TrAC Trends in Analytical Chemistry 27(7), 612–618 (2008)

Bentley, J.L.: Multidimensional binary search trees used for associative searching. Communications of the ACM 18(9), 509–517 (1975)

Cover, T.M.: Estimation by the nearest neighbor rule. IEEE Transactions on Information Theory 14(1), 21–27 (1968)

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Annals of Statistics 32(2), 407–499 (1996)

Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning. Springer series in statistics, Springer, Berlin, second edn. (2009)

Gönen, M., Heller, G.: Concordance probability and discriminatory power in proportional hazards regression. Biometrika 92(4), 965–970 (2005)

Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)

Härmä, H., Peltomaa, R., Pihlasalo, S.: Lanthanide label array method for identification and adulteration of honey and cacao. Analytical Chemistry 87(13), 6451–6454 (2015)

Huang, J., Zhang, T., Metaxas, D.: Learning with structured sparsity. Journal of Machine Learning Research 12, 3371–3412 (2011)

Jakumnee, J., Suteerapataranon, S., Vaneesorn, Y., Grudpan, K.: Determination of cadmium, copper, lead and zinc by flow voltammetric analysis. Analytical Sciences/Supplements 17icas, i399–i401 (2002)

Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97(1–2), 273–324 (1997)

Kot, B., Baranowski, R., Rybak, A.: Analysis of mine waters using x-ray fluorescence spectrometry. Polish Journal of Environmental Studies 9(5), 429–432 (2000)

Naula, P., Airola, A., Salakoski, T., Pahikkala, T.: Multi-label learning under feature extraction budgets. Pattern Recognition Letters 40, 56–65 (2014)

Nowotny, T., Berna, A.Z., Binions, R., Trowell, S.: Optimal feature selection for classifying a large set of chemicals using metal oxide sensors. Sensors and Actuators B: Chemical 187, 471–480 (2013)

Pahikkala, T., Airola, A., Salakoski, T.: Speeding up greedy forward selection for regularized least-squares. In: Draghici, S., Khoshgoftaar, T.M., Palade, V., Pedrycz, W., Wani, M.A., Zhu, X. (eds.) Proceedings of The Ninth International Conference on Machine Learning and Applications (ICMLA 2010). pp. 325–330. IEEE (2010)

Pahikkala, T., Airola, A.: Rlscore: Regularized least-squares learners. Journal of Machine Learning Research 17(221), 1–5 (2016)

Pahikkala, T., Airola, A., Xu, T.C., Liljeberg, P., Tenhunen, H., Salakoski, T.: Parallelized online regularized least-squares for adaptive embedded systems. International Journal of Embedded and Real-Time Communication Systems 3(2), 73–91 (2012a)

Pahikkala, T., Okser, S., Airola, A., Salakoski, T., Aittokallio, T.: Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. Algorithms for Molecular Biology 7(1), 11 (2012b)

Panayappan, R., Venezky, D., Gilfrich, J., Birks, L.: Determination of soluble elements in water by x-ray fluorescence spectrometry after preconcentration with polyvinylpyrrolidone-thionalide. Analytical Chemistry 50(8), 1125–1126 (1978)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

PerkinElmer: Atomic spectroscopy, a guide to selecting the appropriate technique and system (2011), `http://www.perkinelmer.com/Content/relatedmaterials/brochures/bro_worldleaderaaicpmsicpms.pdf`, accessed January 16, 2016

Pesavento, M., Alberti, G., Biesuz, R.: Analytical methods for determination of free metal ion concentration, labile species fraction and metal complexation capacity of environmental waters: a review. Analytica Chimica Acta 631(2), 129–141 (2009)

Pihlasalo, S., Montoya Perez, I., Hollo, N., Hokkanen, E., Pahikkala, T., Härmä, H.: Luminometric label array for quantification and identification of metal ions. Analytical Chemistry 88(10), 5271–5280 (2016)

Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research 11, 2487–2531 (2010)

Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B) 58, 267–288 (1996)

Turku Region Water Ltd.: Water quality table 2016 (2016), `http://www.turunseudunvesi.fi/sites/default/files/halinen_lahtevan_veden_vedenlaadun_koostetaulukko_2016.pdf`, accessed June 16, 2017

Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7(1), 91 (2006)

Wahba, G.: Spline Models for Observational Data, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1990)

Xu, Z.E., Weinberger, K.Q., Chapelle, O.: The greedy miser: Learning under test-time budgets. In: Langford, J., Pineau, J. (eds.) Proceedings of the 29th International Conference on Machine Learning (ICML'12). pp. 1299–1306. icml.cc / Omnipress (2012)

Zhang, T.: Adaptive forward-backward greedy algorithm for learning sparse representations. IEEE Transactions on Information Theory 57(7), 4689–4708 (2011)