Application of machine learning methods for fault detection in Wastewater Treatment Plants (Extended Abstract)

Tatiana Chistiakova Jesùs Zambrano Bengt Carlsson *

* *Uppsala University, Uppsala, Sweden (e-mail: firstname.lastname@it.uu.se).*

Keywords: Fault detection, Machine learning, Classification, Clustering, Wastewater treatment

Wastewater treatment plant (WWTP) is designed to remove chemical, biological waste from water used in agriculture, industry and every day human life. Along with waste products removal, the task is to control various water characteristics. One of such parameters to control in WWTP is the dissolved oxygen (DO) level. The air supplied to the process is proportional to the energy consumption which accounts for 30-50% of the plants total energy use. In order to maintain an effective process and restrict the use of energy, correct DO-measurements are an essential part to analyse and to study in monitoring any plant.

As any automated system, sensors for recording measurements not always can be trustful. Hence, there is a need to control and detect abnormal measurements on time. The goal is to identify faulty washing occurrences based on indication parameters of the DO sensor response to cleaning periods. The indication parameters are the following: amplitude, time constant, rise time and maximum slope.

For the detecting the fault washing occurrences, the popular clustering technique, *k-means* clustering, is applied. The method divides the data set into a number of clusters, based on samples similarity. The assumption is that faulty washing occurrences are grouped together in one of the obtained clusters.

Clustering techniques correspond to unsupervised learning, where no a priori information about the data is given. Hence, the task is to uncover a hidden structure of the data without any known labelling or characteristics. Cluster analysis produces a set of "clusters" (groups), where samples in one cluster are similar to each other and different from the samples of another cluster.

k-means clustering is a widely used method for data analysis. For a finite data set, the k centroids are initialized as random points. Then, the algorithm assigns a sample to the cluster based on a distance measure. Very often, the Euclidean distance D is used, where the distance between two points p and q is

$$D(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + ... + (q_n - p_n)^2} \quad (1)$$

After the convergence, the data set is separated into clusters with different centroids.

Since, no information is evaluable about the data, the essential part is to set the number of clusters properly. For this purpose, the Silhouette index, used for understanding and estimation of clustering results, is applied. The Silhouette index estimates how similar samples in one cluster to samples in another clusters, applying such measure of dissimilarity like distance measures.

As a result, the larger the value of the index, the better quality of a cluster analysis is obtained. The value higher then 0.51 shows a reasonable clustering structure.

In case of studying the data on-line, the classification technique can be used to obtain training data set. Then, the unsupervised learning takes part in the experiment. Several methods are applied for identification of faulty measurements.

To begin with, the naive approach is used when the new measurement is assigned to a class which centroid is the closest.

Then, the well-known k-Nearest Neighbour (k-NN) classifier is performed, where k is the number of neighbours of the given unclassified point. The class is assigned by using majority voting in the k-neighbourhood. The algorithm is simple in its implementation, since it requires no training, but the main drawback is the execution time when the distance measure is computationally complex or the training set is large.

As a more advanced technique, the Support Vector Machine (SVM) technique is used. The SVM classifier mainly maps the data points to a higher and more complex dimensional space so that they are linearly separable. To speed up the computation and handle the potentially high complexity of the mapped space, the kernel trick is used.

The last method to compare is Random Forests, when several decision trees are used to contract a classifier. The new data is studied by every classification tree in the set and, hence, it may be assigned with different classes. The class chosen is the one that is most popular among all trees.

To conclude, the experiments are based on training data obtained from a first part of data analysis when unsupervised leaning is applied. Several machine learning classification techniques are compared in terms of application with fault detection in WWTPs. Also, the experiment is based on obtained training data