Machine Learning for Sensor Analysis

Hyun Kwon [1] And rodney summerscales2
[1] School of Engineering, Andrews University, Berrien Springs, MI
[2] Department of Computing, Andres University, Berrien Springs, MI

ABSTRACT

Machine learning (ML) can be an appropriate approach in many applications including sensor data analysis. ML can provide a way to overcome common problems associated with sensors for low-cost, point-of-care diagnostics. This NSF-funded study proposes a novel approach based on ML algorithms (neural nets, Gaussian Process Regression, among others) to model the electrochemiluminescence (ECL) quenching mechanism of the $[Ru(bpy)_3]^{2+}$/TPrA system by phenolic compounds. The relationships between the concentration of phenolic compounds and their effect on the ECL intensity and current data measured using a mobile phone-based ECL sensor is investigated. ML could provide a robust analysis framework for sensor data with noises and variability. It demonstrates that ML strategies can play a crucial role in chemical or biosensor data analysis, providing a robust model by maximizing all the obtained information and integrating nonlinearity and sensor-to-sensor variations.

MATERIAL AND METHODS

Data Preprocessing

The mobile phone-based ECL sensor simultaneously produces two types of sequential time series data (Figure 1): (1) The ECL light intensity was recorded as a movie file (mp4) by the default camera app, followed by extracting them into image sequences. The average light intensity within the region of interest (ROI) in each frame was calculated using the NIH ImageJ software. (2) The electric current followed by the chronoamperometric voltage application was also recorded by a compact potentiostat in the sensor apparatus. During a 1 sec duration of applied voltage, the first 25 data points of the ECL intensity and 200 data points of the current data were used, as they were the most significant.

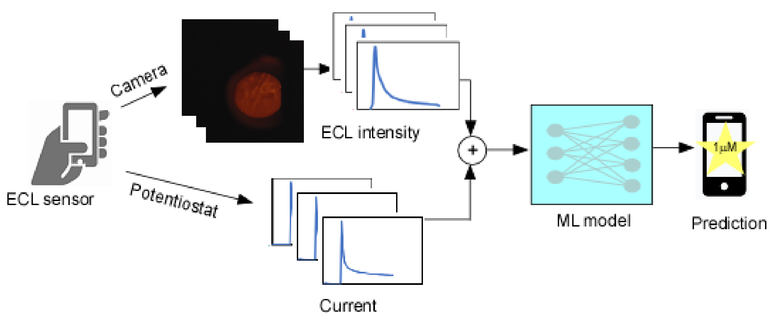

Figure 1. Illustration of entire process of multimodal data collection and prediction process.

Testing Strategy

To evaluate the performance of the concentration prediction model, we used the following ML algorithms: a single, bi-layer and tri-layer neural network, SVR, Boosted Trees, and GPR. Just for comparison, a linear regression method was also used although it would not be the best choice considering non-linear dependencies of the data. The training and test data were split using a stratified shuffle split and a 5-fold cross validation method was used to evaluate the performance more accurately (Figure 2).

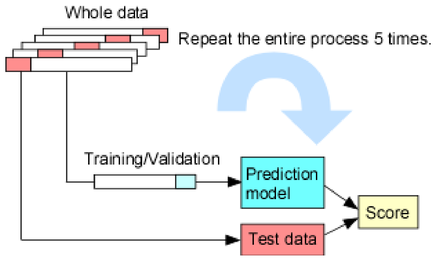

Figure 2. Schematics of 5-fold cross validation where the whole data is split into 5 folds. After the training/validation was completed with the training set (4 folds) in a prediction model, the test set is used to determine the accuracy of the trained model. The entire process is repeated 5 times with each split.

RESULTS & DISCUSSION

The mean absolute error (MAE) values for the prediction of Vanillic and p-Coumaric acids using single, bi and tri-layer neural nets, SVR, Boosted Trees, GPR, and linear regression methods were summarized in Figure 4. First, it is noted that the linear regression method, equivalent to a traditional calibration curve, performed significantly worse (higher MAE) than the ML models for both Vanillic and p-Coumaric acid data, indicating the nonlinear dependencies of both the intensity and current signals to the concentration of the phenolic compounds. The ML models were effective for comprehending the nonlinearity of the sensor signals to the concentration. Second, it is observed that ML using multimodal data (combined ECL intensity and current data) was effective in achieving better prediction performances. For instance, for Vanillic acid, a significantly reduced MAE was achieved using multimodal data, indicating ML can identify relationships between the intensity and current in order to infer the concentration of the phenolic compounds.

Two notable ML models are tri-layer neural net and GPR, which produced out-standing results consistently. A common practice for providing input variables (predictors) is using features extracted from sensorgrams. In this case, the number of predictors is seven for each intensity and current signal, fourteen when combined. We have also used the preprocessed time series data as predictors, which is close to the raw data directly from experiments. The intensity had 25 and the current had 200 predictors (due to the different sampling rates). Using time series values as predictors saves significant preprocessing and feature extraction efforts. Feature engineering, developing informative features for ML algorithms, is often challenging.

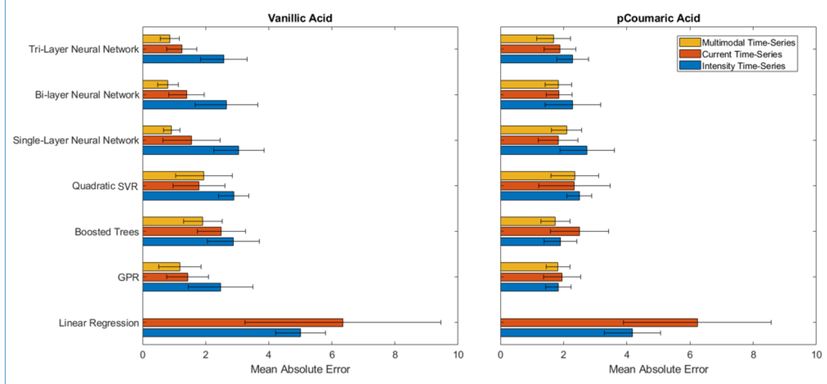

Figure 4. MAE test results from various ML models that were trained from the time series of multimodal (combined intensity and current), current alone, and intensity alone for Vanillic and p-Coumaric acids. The ML models were trained from 70 to 80 measurement experiments in the range of 0.1–30 μM and 1–50 μM for Vanillic acid and p-Coumaric acid, respectively. The error bars represent the standard deviation from the 5-fold cross validation method.

The proposed ML models successfully predict concentration given data collected from ECL sensors. Considering the complex nature of electrochemical reactions in the ECL quenching mechanism by phenolic compounds, it is remarkable that the ML models can achieve this from sensor time series values without extensive preprocessing and feature extraction. Developing a mechanistic or first-principle model for prediction purposes and which also explains electrochemical reactions and mass transport mechanisms on the circular electrodes is complex and time consuming. Even with such a model, there may be no guarantee of its effectiveness in a data analysis pipeline. This study demonstrates that ML models provide accurate predictions of the concentration of phenolic compounds and can account for sensor-to-sensor variations.

CONCLUSION

The low-cost, mobile phone-based ECL sensor generated nonlinear, multimodal data with considerable variability due to sensor-to-sensor variations and environmental fluctuations. In contrast to the traditional calibration approach, the ML models, such as tri-layer neural net or Boosted Trees, carried out effective regression tasks for detection purposes by learning higher pat-terns from the multimodal data. The results demonstrated that the ML models could provide a robust analysis framework for sensor data with noises and variability with-out extensive preprocessing. The ML analysis can compensate for the deficiencies of less stringent, simple, affordable device settings through powerful learning algorithms and thus, accelerate the implementation of low-cost sensors in a wide range of practical situations, such as the detection of phenolic compounds on-site and their monitoring in industrial environments.

ACKNOWLEDGEMENTS: