DATA DRIVEN SOFT SENSOR FOR CONDITION MONITORING OF SAMPLE HANDLING SYSTEM (SHS)

Abhilash Pani, Jinendra Gugaliya and Mekapati Srinivas

Industrial Automation Technology Centre, ABB, Bangalore, India

ABSTRACT

Gas sample is conditioned using sample handling system (SHS) to remove particulate matter and moisture content before sending it through Continuous Emission Monitoring (CEM) devices. The performance of SHS plays a crucial role in reliable operation of CEMs and therefore, sensor-based condition monitoring systems (CMSs) have been developed for SHSs. As sensor failures impact performance of CMSs, a data driven soft-sensor approach is proposed to improve robustness of CMSs in presence of single sensor failure. The proposed approach uses data of available sensors to estimate true value of a faulty sensor which can be further utilized by CMSs. The proposed approach compares multiple methods and uses support vector regression for development of soft sensors. The paper also considers practical challenges in building those models. Further, the proposed approach is tested on industrial data and the results show that the soft sensor values are in close match with the actual ones.

KEYWORDS

Sample Handling System, Soft-Sensor, Variance Inflation Factor (VIF), Local Outlier Factor (LOF), Support Vector Regression.

1. INTRODUCTION

Adverse impacts of rapid industrialization on world's environment are acknowledged worldwide which are mostly irreversible. Hence, various government agencies along with industries have started monitoring emissions to control associated environmental pollution. Continuous emission monitoring (CEM) devices are used across industries for monitoring real-time pollutant content in flue gases [1] [2] . As governments across the globe becomes more vigilant and stringent on emission norms, reliability and availability of CEM systems have become very crucial. Reliability of CEM systems are dependent not only on CEM devices but also on associated sample handling systems (SHSs). As reported in [3] majority of failures in CEM systems are due to issues in SHSs. Therefore, manufactures have started offering sensors for condition monitoring of SHSs, which helps in improving reliability of these systems and hence reliability of overall CEM systems.

Performance of any condition monitoring system depends on sensors and is adversely impacted by sensor failures. Therefore, many studies have discussed methods for sensor fault detection and isolation. In [4] autocorrelation is used to detect sensor failure of pitot static system in airplanes. Spectral clustering technique based faulty sensor detection and deletion from wireless sensor network is proposed in [5] . [6] provides a detailed review on sensor fault detection methods and report that 40% of literature on sensor fault detection are based on either principal component

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 307-320, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101423

analysis (PCA) or artificial neural network (ANN). [7] is one of the earliest and most cited paper on using PCA for sensor fault detection. In [8] feed forward neural network and Locally weighted regression are proposed for sensor fault detection in Predictive Emission Monitoring System.(PEMS) unlike traditional CEMs, pollutant concentration is estimated using process data and parameters instead of measuring them directly. Once faulty sensor is detected and isolated using fault detection technique, data driven soft sensor model is used to estimate the true value of the faulty sensor. These estimates can be used in place of faulty hardware sensor measurement which adds fault resilience characteristics to condition monitoring systems [9] . Recently [10] has proposed a deep learning-based vision sensing applied to printing quality control. Online adaptive ensemble PLS approach is proposed for a chemical process in [11] A Gaussian probabilistic regression approach [12] is proposed to develop soft sensor. A hierarchical clustering method is proposed in [13] .

However, in literature there is not enough material which provides a detailed framework for development of data driven soft sensor for SHS. This is essential as there are practical constraints which are important to be considered during development of soft sensors.

This work focuses on a framework to develop data driven soft sensor for condition monitoring of SHSs. There are few practical assumptions/constraints which are considered during development of the framework. They are mostly related to SHS system and are discussed in the upcoming sections. Rest of the paper is organized as follows. Section 2 introduces Sample Handling System, Section 3 discusses the objective of the work, Section 4 details on the proposed framework, Section 5 provides the results with an use case and Section 6 concludes the study.

2. SAMPLE HANDLING SYSTEM

The main objective of SHSs is to (1) Provide a path for sample collection (2) Transport collected sample without contamination (3) Remove particulate matter and moisture present in the gas sample (4) Maintain desired temperature and regulated flow of gas sample to CEM device.

There are multiple stages to ensure above objectives are achieved. Figure 1 provides a block diagram of SHSs. Each stage in Figure 1 will have one or more components and to monitor functioning of these components there are multiple sensors placed across SHS. The SHS considered in this analysis has 2 temperature sensors, 3 pressure sensors and one flow sensor.

The first stage of SHS consists of sample probe and filter components which are responsible for collecting sample gas and to filter particulate matter present in it. Normally this stage is placed close to exhaust stack and away from remaining stages of SHS. The distance between first stage and remaining stages varies from plant to plant and in some cases can reach values close to 500 meters. The second stage of sample processing consists of temperature treatment which ensures that temperature of collected sample does not drop below certain threshold value to prevent condensation of available moisture. The third stage is responsible for removing moisture by cooling the incoming gas sample. This stage also removes the collected condensate from the process. Fourth stage contains sample pump which is the heart of the SHS. Sample pump produces a pressure difference to ensure enough flow of sample gas to CEM devices. Control valves present in fourth stage regulate the flow as per design specification.

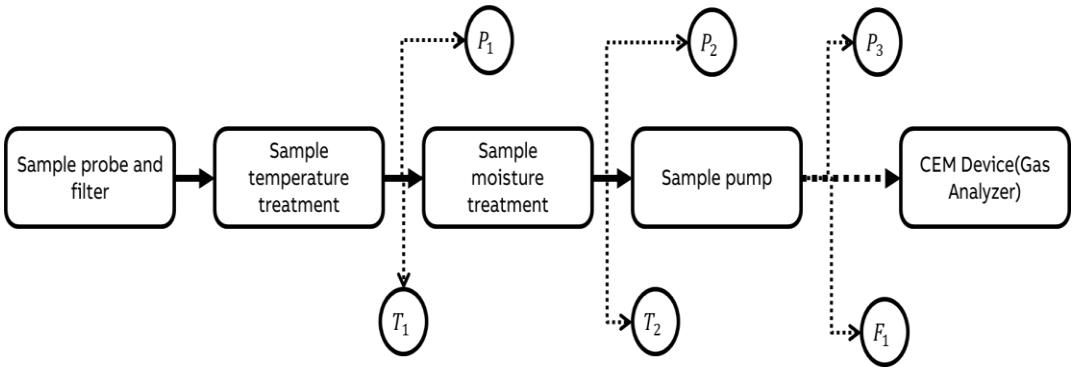


Figure 1: Block Diagram of Sample Handling System

3. OBJECTIVE AND ASSUMPTIONS

The objective of this work is to build a framework/pipeline to develop soft sensors for SHS such that each soft sensor can model measurement of a faulty hardware sensor using measurements from other hardware sensors. In order to do so there are few assumptions which were made considering practical experience.

Assumption 1. The framework was developed for single sensor failure only, which means developed soft sensors will only work if there are single sensor failures. This assumption was considered as simultaneous failure of multiple sensor are rare in field. Secondly the time between two successive hardware sensor failures is large enough to schedule a planned shutdown of process to replace failed sensor.

Assumption 2: Availability of training data will be less (~4000 samples). As developed framework utilizes machine learning approach, historical data from all hardware sensors are required for training. For a new installation getting a large amount of training data is not feasible. Therefore, in this study we have not considered machine learning algorithms which need large volume of data for its training.

Assumption 3: The available training data is collected during normal operation of SHS. For a new installation with all testing done, it is extremely rare to encounter fault/failure and therefore, this assumption should be validated before developing soft sensor.

4. FRAMEWORK

The proposed framework for soft sensor development can be divided into two major modules.

1. Data Pre-processing module
2. Machine learning algorithm evaluation and selection module.

4.1. Data Pre-Processing Module

This is the first module in the framework which ensures quality of data for modeling of soft sensor. This module performs 5 data preprocessing steps, and the details of each step are given in the following.

4.1.1. Missing Value Imputation

Missing observations are common in any data and there are many imputation methods available in literature [14] . Each method has its own advantage and disadvantage and the best imputation method mainly depends on amount of missing data and its type. In this work we have considered time series data from SHS and hence central tendency-based imputation methods (mean/median imputation methods) are not suitable. Imputation by last observation carried forward, imputation by next observation carried backward and imputation by interpolation are three popular methods used for imputation in time series analysis. In this study imputation by interpolation is considered as it considers both previous and next observation value for imputation.

4.1.2. Removal of Off Condition and Outlier

In the training data there are off conditions where SHS is offline. These samples with off condition should be removed before proceeding for further analysis as these samples may impact soft sensor models adversely. The off condition can be checked using sensor measurement. In this study, SHS is considered off-line if all pressure measurements are 0.

Given a sample from an unknown population a point is labeled as outlier if it is located away from majority of the samples. This is known as distance-based outlier definition. According to density-based outlier definition, a point is labeled as outlier if the point is present in a low-density region in multidimensional feature. There are many approaches for outlier detection and removal. In this work we have considered a two-stage outlier removal process in which stage 1 removes distance-based outliers whereas stage 2 removes density-based outliers.

In stage1, quartile-based univariate thresholds for each variable are calculated as in the following

$$\text{Upper Threshold} = Q_3 + 1.5 \times IQR$$
$$\text{Lower Threshold} = Q_1 - 1.5 \times IQR$$

Where $Q_1$and Q_3represents first and third quartiles and $IQR$is the inter quartile range calculated as $IQR = Q_3 - Q_1$.This is a simple distance based univariate approach and is performed first to remove outliers which are far away from majority of population.

In stage 2, density-based local outliers are removed using Local outlier factor (LOF). LOF assigns an outlier score to each sample based on relative density of that sample with respect to its K nearest neighbors. More details on LOF can be found in[15] .

4.1.3. Removal of Features with Multicollinearity

Multicollinearity is a phenomenon in which one or more independent variables can be expressed as a linear combination of other independent variables. In the presence of multi collinearity influence of independent variables on target variable cannot be estimated accurately. Therefore, interpretation of trained models becomes difficult.

Correlation among independent variables can be calculated using Pearson correlation or Spearman correlation coefficient. These correlation coefficients calculate correlation among two independent variables at a time, which is a major limitation. Therefore, in proposed framework, Variable inflation factor (VIF) score is used to identify and remove correlated variables. VIF score for an independent variable is calculated by regressing it against every other independent variable in the model according to the below equation.

$$VIF_i = {}^{1}/_{(1-R_i^2)}$$

Where VIF_i is the VIF score for i^{th} independent variable and R_i^2 is the coefficient of determination (R_squared value) obtained by regressing i^{th} independent variable against every other independent variable. A VIF score of 1 indicates no correlation. A VIF score of more than 10 is considered as extremely correlated and corresponding variable /feature should be dropped. However, dropping all variables with VIF score more than 10 at once is not a good strategy. For example, let's take a regression with 4 independent variables $[X_1, X_2, X_3, X_4]$ and assume that X_1, X_3 and X_4 not correlated. However,X_2 can be expressed as $3X_3 + 4X_4$. By rearranging this equation we can show that there is multicollinearity among $[X_2, X_3, X_4]$. Therefore, for this set up, VIF score for X_1will be minimum and will be higher for $[X_2, X_3, X_4]$. Let's assume that VIF score for X_2, X_3 and X_4 is greater than 10 and if all variables with VIF >10 are dropped at once loss of uncorrelated variables may happen (in this case X_3 and X_4). This is not an efficient way. Therefore in this framework an iterative removal of features based on VIF score is proposed and detailed steps are shown in

Figure 2.

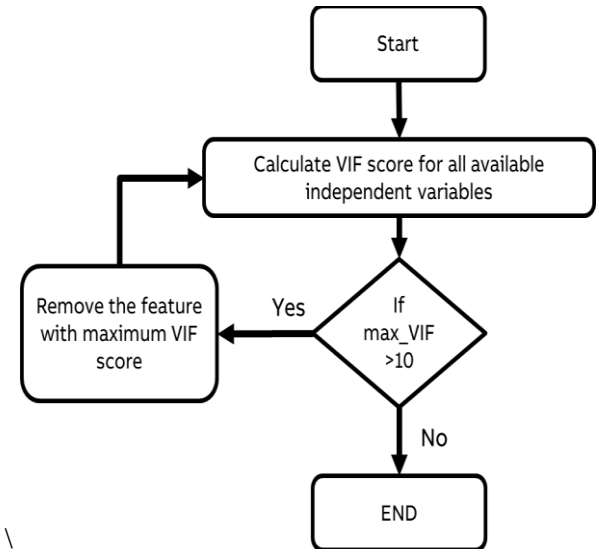


Figure 2 Flow chart of Iterative feature removal method

4.1.4. Time based Data Split

It is a standard practice to divide data into three sets namely train, validation and test datasets. Popularly this is done at random with each sample having uniform probability. This is not a suitable strategy for splitting time series data. In case of time series there is an inherent temporal dependency and hence the test accuracy obtained by random splitting will be misleading. There for in this framework, time-based data splitting is used. A visual representation of random and time series splitting is provided in Figure 3.

(a) (b)

Figure 3: (a) Random splitting and (b) Time-based splitting

4.1.5. Data Normalization

Data normalization is the process of transforming each independent variable such that transformed variables will have 0 mean and 1 standard deviation and normalization step should follow data splitting step. The sequence of last two preprocessing steps is important as normalization of test data should happen based on training data distribution otherwise data leaking issue will arise.

4.2. Machine Learning Algorithm Evaluation and Selection Module

Considering assumption 2 (in Section 3), we have considered 5 algorithms in this framework. Details of these algorithms can be found in [16] and [17] .

1. Linear regression
2. K nearest neighbour
3. Decision Tree
4. Random Forest
5. Support vector regression

There are other machine learning algorithms like ensemble models (GBDT, stacking) and Neural network which are proven accurate for learning complex relation between independent and dependent variable. However, due to their higher flexibility they are prone to overfitting. Considering low volume of training data overfitting issue will worsen. Therefore, these algorithms are not considered in the framework.

In this module hyper parameter tuning for each of the algorithms is done using grid search. In order to evaluate these algorithms, mean square error is considered in this framework. After evaluation, the best model is considered for modeling of soft sensor.

5. RESULTS AND DISCUSSION

In this section results of soft sensor models built using proposed framework are presented. Process data from one of the CEM system installation in India was used for evaluation of the proposed approach. One month of process data with 4,320 data points with 6 features is used for building regression models for soft sensor. The feature values are standardized to 0 mean with 1 standard deviation. In this work temperature measurements are represented as T_1 and T_2. Similarly, pressure measurements are represented as P_1 , P_2 and P_3 and flow measurement is represented as F_1.

After removal of missing values and off conditions from available dataset, two stage outlier removal is performed. As discussed, earlier distance-based outliers are removed in the first stage followed by removal of density-based outliers using LOF scores in second stage. In order to visualize identified outliers, multidimensional feature space is embedded into two-dimensional space using t-distributed Stochastic Neighbour Embedding (t-SNE). In Figure 4, scatter plot of dataset with identified outliers is shown. In this plot inliers/normal data points, outliers identified using IQR method (distance based) and outliers identified by LOF score (density based) are represented with circular, star and square markers respectively. Number of data point belonging to normal, distance based outlier and density based outlier are provided in Table 1. Distance based outliers are located towards outer edges of the scatter plot, whereas density-based outliers are present towards inner side of the plot.

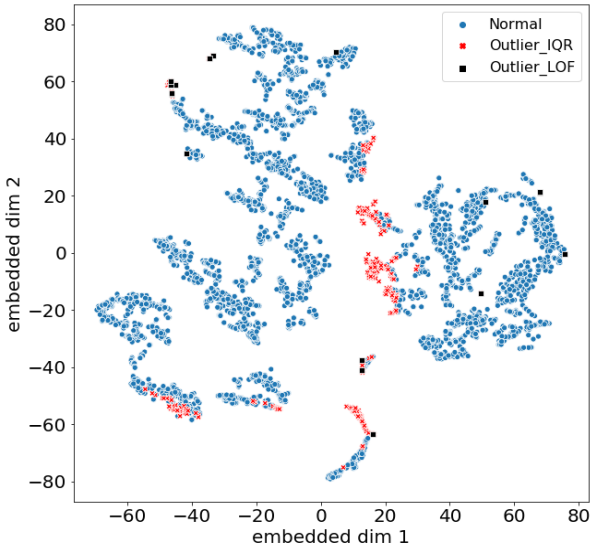


Figure 4: Outliers in the given data set

Table 1: Number of normal data points and outliers

Point type	**Number of samples**
Normal	3813
Distance based Outliers	312
Density based Outliers	21

Next step in data pre-processing is the removal of features with multicollinearity using a recursive method. In oder to showcase results of proposed method let's consider the case of modeling of soft sensor for T_2. where, remaining 5 measurements are considered as predictors and multicolinearity is evaluated for these 5 predictors. Feature-wise log VIF scores are plotted in Figure 5 and absolute VIF scores are provided in Table 2. VIF score of 10 is considered as threshold for feature elimination. From the plot it is evident that P_1 and P_2 have VIF score more than the threshold.

Using recursive feature elemination method first P_1 is removed from predictor list as it has highest VIF score and VIF scores for remaining 4 predictors are evaluated again. VIF scores in second itteration arepresented in Table 2. As evident from this table,after droping P_1, VIF scores for remaining 4 predictors including P_2 are less than 10, which indicates no sever multi colinearity. Therefore, as discussed earlier , droping P_1 and P_2 at once based on threshold is not a good statergy as multicolinearity in P_2 can be eliminated by droping P_1 only.

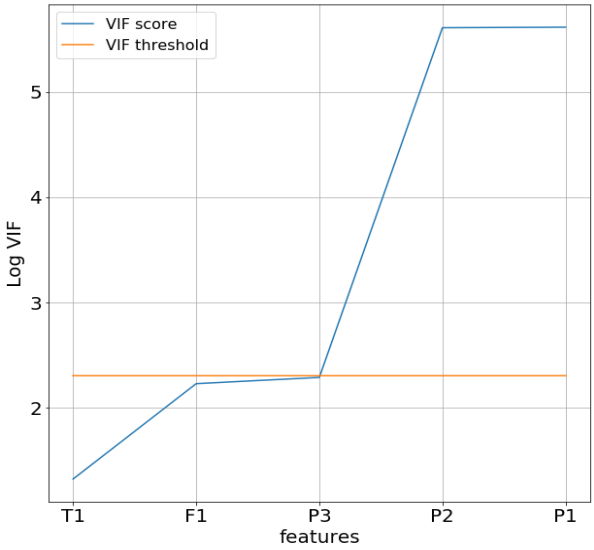


Figure 5: Log VIF Scores of all independent features

Table 2: VIF Scores of Different Features

variables	**VIF score**	
	Iteration 1	**Iteration 2**
T_1	3.75682	1.298446
F_1	9.296621	3.734248
P_3	9.861192	6.246676
P_2	273.2662	7.746631
P_1	274.688	---

To demonstrate impact of features having sever multicollinearity on regression model, modelling of soft sensor for P_3 is considered. Two different models were trained using decision tree algorithm. All 5 features are considered for training of model 1 whereas, feature P_1 is dropped from feature list while training model 2. Model 1 identified $P_1$and P_2 as top two important features for prediction of P_3. However, order of feature importance and their corresponding values change drastically when training datset was changed slightly. This makes interpretation of feature iportance difficult in preesence of multicolinearity . P_2 and F_1 are identified as top two important features by modle 2. Order and value of feature importance are consistant compared to model 1 which makes impterpretation easier. This problem becomes even more significant for systems with large number of features and therefore it is a good practice to remove feature with sever multicolinearity.

After removal of features with multi colinearity first 2,860 (~75%) samples are considered as training data set and remaining as test dataset. These datasets are used to train and evaluate soft sensor models developed using 5 different machine learning algorithms namely linear regression(LR), K nearest neighbour (KNN), support vector regression(SVR) , decision tree(DT) and random forest(RF). Hyper parameter tunning for each algorithm is perfomred by further splitting training data into train and cross validation dataset. Table 3 provides tuned parameter vaues and comparison of aforementioned algorithms is performed using mean square error.

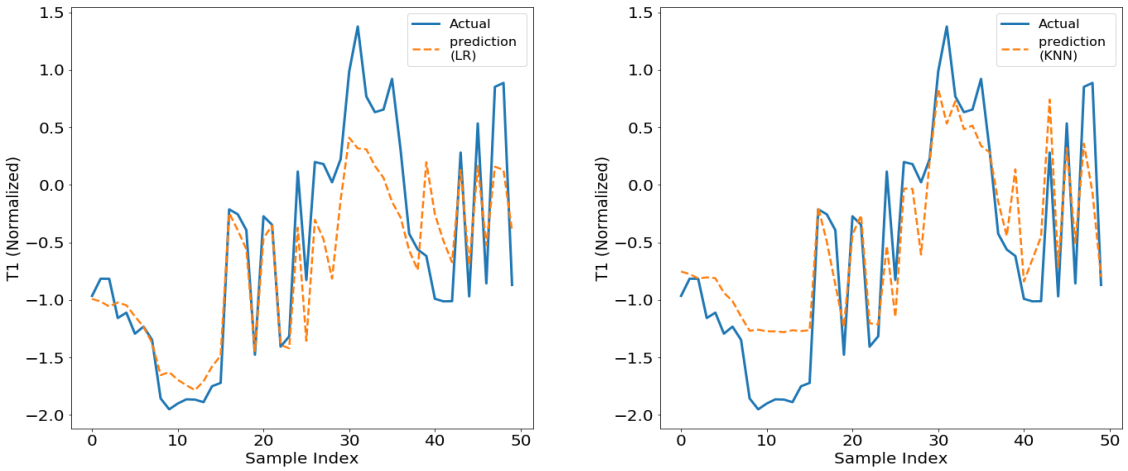
Table 3 Parameter values after hyperparameter tuning

Algorithm	**Parameter values**
Linear regression	L2 Regularization parameter = 0.1
K nearest neighbour	Number of nearest neighbour = 50
Decision tree	Maximun depth = 5
Random forest	Maximum Depth = 50, Number of estimator = 500
Support vector regression	Regularization parameter(C) = 10, Kernel = "RBF", Kernel coefficient (gamma) = 0.2

To compare above mentioned algorithms, modeling of soft sensor for T_1 is considered and the calculated mean sqaure error values for both train and test datasets are given in Table 4. From this table it is evident that SVR has minimum test MSE followed by RF. However, the difference in train and test MSE for model obtained by RF algorithm has higher varience issue. This issue can be avoided by increasing number of base estimators in RF given higher volume of training data. Due to the constraint of low volume of training data (assumption 2) ensembling algorithms and neural networks are not used for this application. Plot of actual and predicted values by various algorithms is shown in Figure 6 for visual comparison. The prediction by SVR model is very close to actual values of T_1 followed by that of RF and DT. Predicted values of LR and KNN models could not capture peak patterns in T_1 which are captured by other three algorithms. From the above comparative analysis, SVR with RBF (radial basis function) kernel is selected for modeling of soft sensor in SHSs.

Table 4: MSE values for different ML Algorithms for $\boldsymbol{T_1}$ Soft Sensor Model

Algorithms	**Linear Regression**	**KNN**	**SVR**	**Decision Tree**	**Random forest**
Train MSE	0.403	0.214	0.134	0.310	0.111
Test MSE	0.307	0.306	0.134	0.289	0.236



(a) (b)

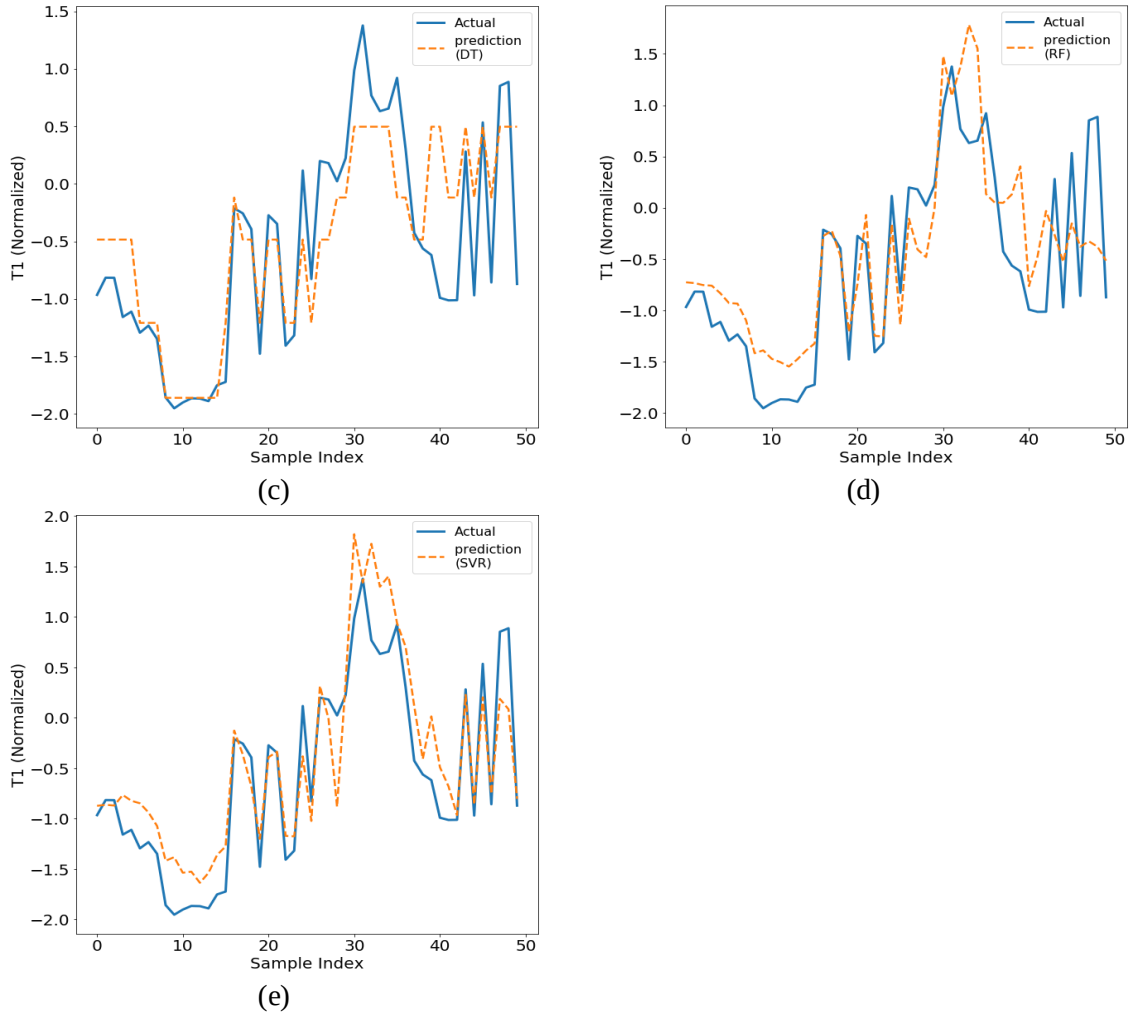


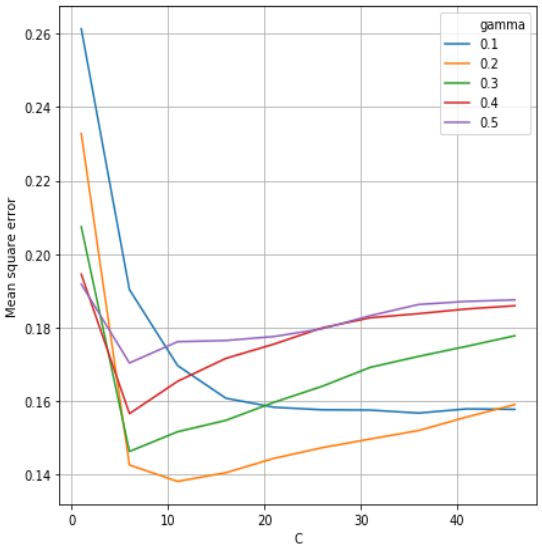
Figure 6: Actual and Predicted Values for $\boldsymbol{T_1}$ Soft Sensor (a) Linear regression (b) K nearest Neighbour (c) Decision Tree (d) Random Forest (e) SVR

Performance of SVR with RBF kernel is superior as it can learn complex nonlinear function by projecting data to higher dimension using Kernel trick. T is advantage becomes more significant when amount of training data is limited. Impact of tuning parameters on performance of SVR is shown in Figure 7. Two parameters of SVR with RBF kernel are considered in this analysis. Parameter C is the regularization parameter of SVR, and strength of the regularization is inversely proportional to C. Parameter gamma represents spread of radial basis function. A range of values for C and gamma are selected and a grid search approach is used for parameter tuning. From Figure 7 it is evident that minimum MSE on cross validation data was obtained for C= 10 and gamma = 0.2. Hence a soft sensor model for T_2 using SVR with RBF kernel, C= 10 and gamma = 0.2 is developed and plot of actual and predicted value by this model is shown in Figure 8. From this figure it can be observed that predicted values could capture overall trend, however model is not able to capture extremely peak patterns in actual data.

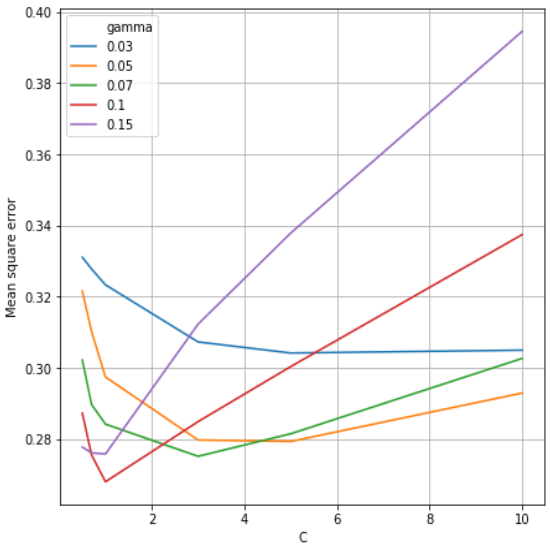
Similar approach is adapted for modeling soft sensors for other measurements/physical sensors and obtained test and train mean square error along with R square values are provided in Table 5.

Table 5: R Square and MSE values for Different Sensors

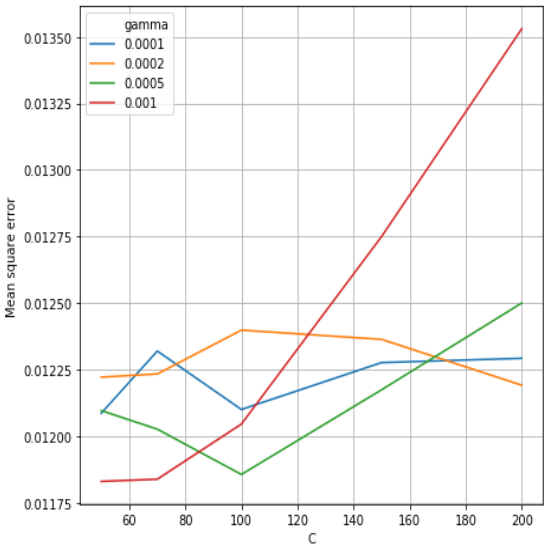
Physical sensor	**Test MSE**	**Train MSE**	**Test R square**	**Train R square**
T_1	0.23	0.13	0.83	0.89
T_2	0.15	0.20	0.78	0.80
P_1	0.011	0.008	0.83	0.93
P_2	0.006	0.007	0.96	0.99
P_3	0.002	0.001	0.85	0.90
F_1	0.04	0.02	0.746	0.83



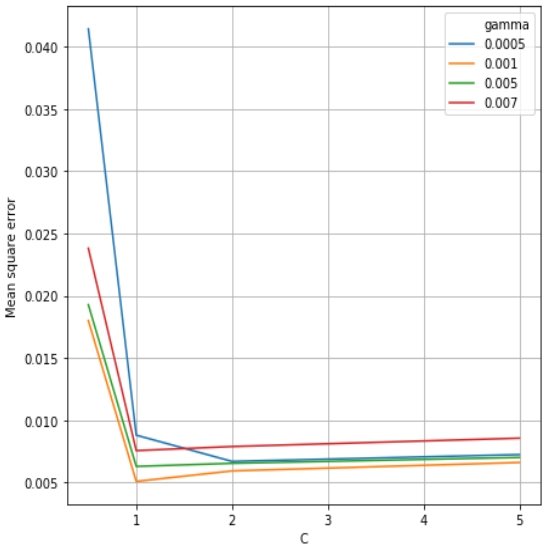
T_1



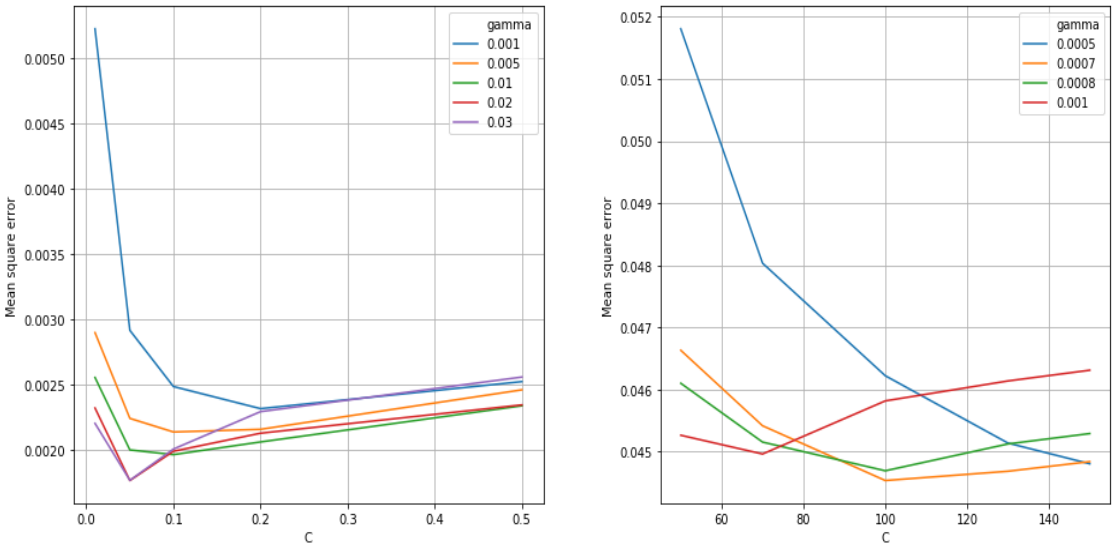
T_2



P_1

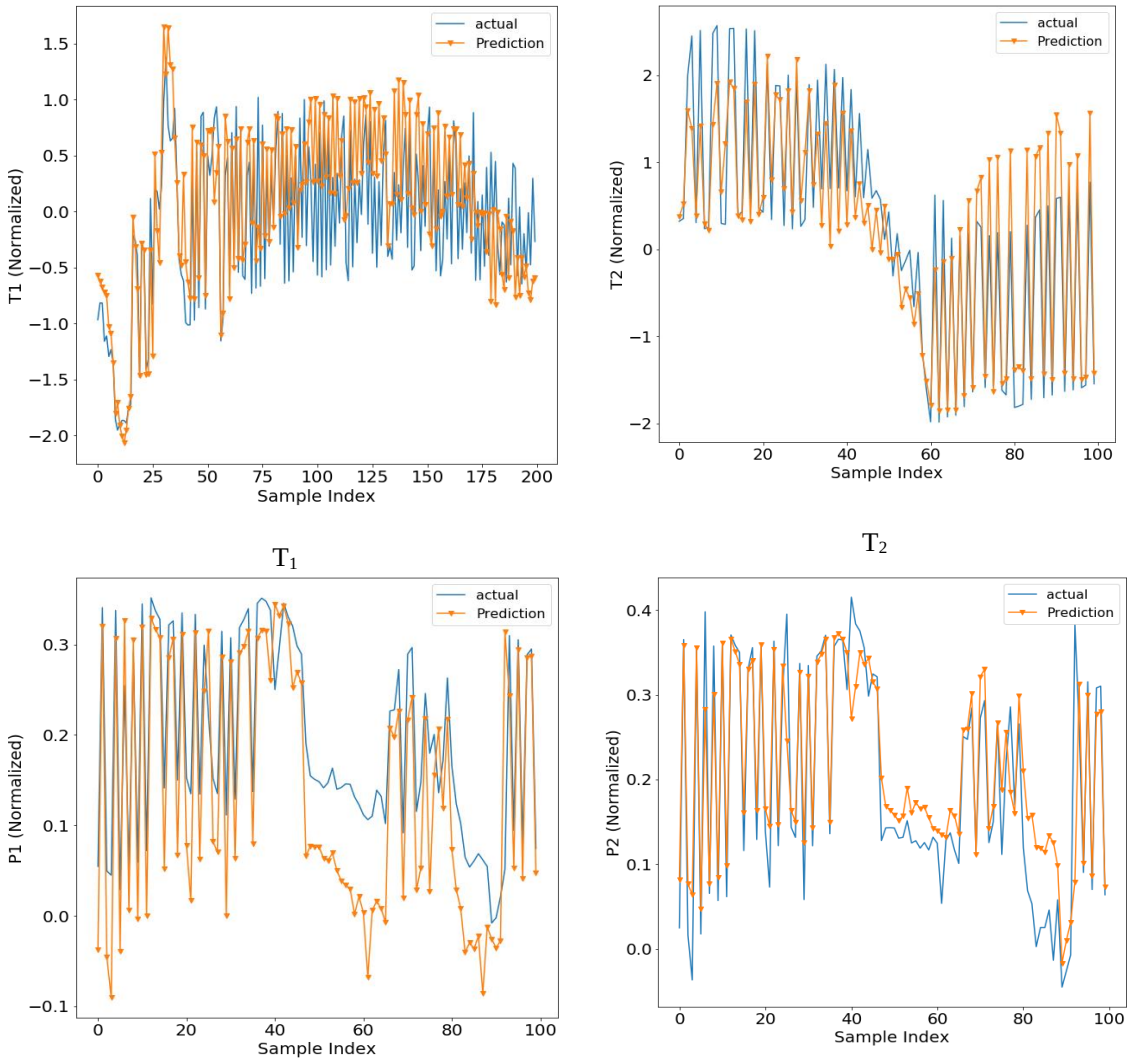


P_2



P_3 F_1

Figure 7: Plot of MSE for various Gamma and C values of SVR for 6 Soft Sensor Models



T_1 T_2

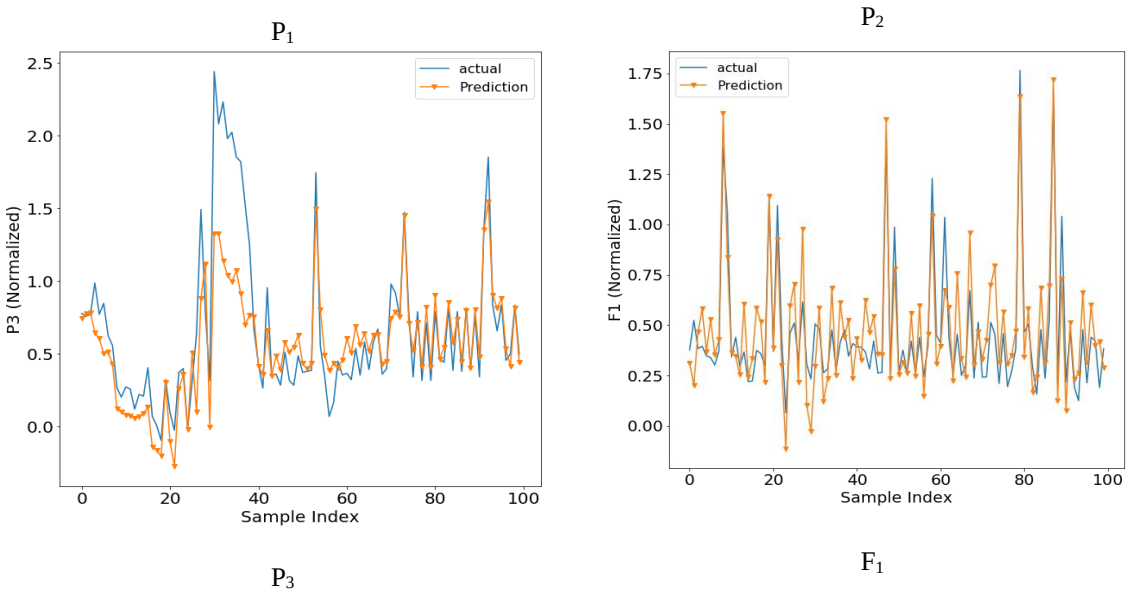


Figure 8: Actual and Predicted Values for 6 Sensors using SVR

6. CONCLUSIONS

In this study a framework is proposed for developing data driven soft sensor for sample handling system in CEMs. The framework consists of two modules: (i) Data Preprocessing module (ii) Machine learning algorithm evaluation and selection module. In module (ii), 5 machine learning algorithms which includes Linear Regression, KNN, SVR(RBF), Decision Tree and Random Forest are evaluated on industrial data that consists of 6 SHS measurements. From the comparison SVR is found to be better than other methods in predicting the values for all the 6 SHS measurements. The future work would involve exploration of Deep ML approaches and compare their performance against the proposed approach when data availability is not a constraint.

REFERENCES

[1] J. Jahnke, Continuous Emission Monitoring. Wiley, 2000.
[2] E. Arioni, N. Bonavita and M. Paco, “Keeping an eye on emissions,” Hydrocarbon Engineering Magazine, vol. 18, no. 10, pp. 43–49, October 2013.
[3] Y. Yang, X. Zhang, Z. Zhao, G. Wang, Y. He, Y. Wu and J. Li, "Applying Reliability Centered Maintenance (RCM) to Sampling Subsystem in Continuous Emission Monitoring System" IEEE Access, vol. 8, pp. 55054-55062, 2020.
[4] Swischuk, Renee C. and Douglas L. Allaire. “A Machine Learning Approach to Aircraft Sensor Error Detection and Correction.” Journal of Computing and Information Science in Engineering, vol. 19, pp. 1-19, 2019.
[5] A. M. T. Nasser and V. P. Pawar, "Machine learning approach for sensors validation and clustering," 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, pp. 370-375, 2015.
[6] H.Y. Teh, A. W., Kempa-Liehrand K. Wang. “Sensor data quality: a systematic review.” Journal of Big Data, vol. 7, pp. 1-49, 2020.
[7] R. Dunia, S. J. Qin, T. F. Edgar and T.J. McAvoy, “Use of principal component analysis for sensor fault identification” Computers & Chemical Engineering, vol. 20, pp. S713-S718, 1996.

[8] G. Ciarlo, E. Bonica, B. Bosio and N. Bonavita, “Assessment and Testing of Sensor Validation Algorithms for Environmental Monitoring Applications”, Chemical Engineering Transactions, vol. 57, pp. 331-336, Mar. 2017.
[9] D. Angelosante, M. Guerriero, G. Ciarlo and N. Bonavita, "A Sensor Fault-Resilient Framework for Predictive Emission Monitoring Systems," 21st International Conference on Information Fusion (FUSION), Cambridge, pp. 557-564, 2018.
[10] Villalba-Diez, J.; Schmidt, D.; Gevers, R.; Ordieres-Meré, J.; Buchwitz, M.; Wellbrock, W. Deep Learning for Industrial Computer Vision Quality Control in the Printing Industry 4.0. Sensors 2019, 19, 3987.
[11] Cang, W.; Yang, H. Adaptive soft sensor method based on online selective ensemble of partial least squares for quality prediction of chemical process. Asia-Pac. J. Chem. Eng. 2019, 14, 2346.
[12] Xiong, W.; Shi, X. Soft sensor modeling with a selective updating strategy for Gaussian process regression based on probabilistic principle component analysis. J. Frankl. Inst. 2018, 355, 5336–5349.
[13] Yu, W. A mathematical morphology based method for hierarchical clustering analysis of spatial points on street networks. Appl. Soft Comput. 2019, 85, 105785.
[14] W. Young, G. Weckman and W. Holland, “A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits.“ , Theoretical Issues in Ergonomics Science, vol. 12(1), pp. 5–43, 2011.
[15] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and Jorg Sander, "LOF: Identifying Density-Based Local Outliers", Proc. of the 2000 ACM SIGMOD on Management of Data, pp. 93-104, 2000.
[16] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press, 2014.
[17] H. Drucker, C. Burges,L. Kaufman, A. Smola and V. Vapnik, "Support Vector Regression Machines", Proceedings of the 9th International Conference on Neural Information Processing Systems (NIPS'96). MIT Press, Cambridge, MA, USA, 155–161.

AUTHORS

Abhilash Pani is currently working as scientist ABB Industrial Automation Technology Centre Bangalore and his areas of interests are applied machine learning for industrial analytics and condition monitoring of industrial assets.

Jinendra K Gugaliya is working as Principal Scientist at ABB Industrial Automation Technology Centre Bangalore and his areas of interests are applied machine learning for industrial analytics, reinforcement learning for industrial process controls, model predictive control, and advanced optimization.

Mekapati Srinivas is working as Principal Scientist at ABB Industrial Automation Technology Centre Bangalore and his areas of interests are process modelling, simulation and optimization.