Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

Product Attribute Forecast: Adaptive Model Selection Using Real-Time Machine Learning

Elif Seyma Bayrak[1], Tony Wang[1], Aditya Tulsyan[2], Myra Coufal[2], Cenk Undey[1]

[1]Amgen Inc. Process Development, Thousand Oaks, CA 91320 USA (Tel: 805-447-2159; e-mail: ebayrak@amgen.com, tonyw@amgen.com, cundey@amgen.com)
[2]Amgen Inc. Process Development, Cambridge, MA 02142 USA (atulsyan@amgen.com, mcoufal@amgen.com)}

Abstract: A real-time machine learning framework is developed to forecast product concentration in mammalian cell culture bioreactors. In real-time, the framework evaluates several machine learning algorithms and chooses the most representative algorithm based on current dynamics of the system. Data from multiple sources is combined and only subset of features are fed to the model based on a pre-selection criteria. The model performance is tested using two small-scale bioreactors run. The performance improved towards the end of the process with accumulating data and results for 1 day ahead prediction is presented.

Keywords: Real-time machine learning, product quality attributes forecast, adaptive model selection.

1. INTRODUCTION

The advances in sensor technology, computational resources and artificial intelligence are progressing at a rapid pace. Extensive size of data is collected during process characterization studies in process development (PD) from numerous bench-scale bioreactor experiments but only limited part of this information is analysed in detail beyond the purpose of the specific experiments. With the advances in process analytical technologies, we have been generating better and more informative data and unfortunately often times these rich information are typically not used and just archived.

In instances where this potentially powerful data is analysed for predictions of desired attributes, the modeling expert needs to select a type of model to use for the dataset. The complexity involved in the production of biologics makes it difficult to choose one perfect approach that can capture all the dynamic relations. Distinctive phases of cell culture and different nature of product quality attributes introduce even more challenges to use one modelling approach since prediction power is quite interchangeable. Machine learning (ML) is such a concept where computational models learn from the data, evolve over time without needing human intervention (Stefik, 1985).

There are various ML algorithms capable of identifying different structures of correlation in the data and recently are being used more in biopharmaceutical industry. In one study, performance based models were generated using different statistical algorithms including support vector machines (SVM), partial least square regression (PLSR), random forest (RF) for final product concentration and quality attributes prediction. The study concluded that prediction improved later in the culture and different algorithms performed better for different response variables (Schmidberger et al., 2015). In Le's study, SVM and PLSR were used to predict the final antibody concentration and the final lactate concentration (Le et al., 2012). It showed that both the final antibody concentration and the final lactate level were predicted more accurately when data from the early stages of the production process was used. In another study, five supervised machine learning algorithms including SVM, RF, naïve Bayes classifier (NBC), K nearest neighbor (KNN), and PLSR were used to predict deamination of proteins and the metrics to compare these algorithms were discussed (Jia and Sun, 2017). On downstream side of the process, Agarwal et al. applied artificial neural networks (ANN) approach to predict depth filter loading capacity for clarification of the monoclonal antibody from the cell culture (Agarwal et al., 2016). These studies showed the varying performance of ML algorithms depending on the dynamics of the system and nature of the attributes to be predicted, however real-time learning and adaption of modelling algorithm to these changing dynamics were not studied.

In this work, we are proposing an alternative methodology that starts learning with *the first data generated* while combining information from different data sources from the very first experiment in process development that could potentially be implemented in manufacturing. The proposed algorithm does not assume any prior modeling approach, instead it evaluates and ranks different approaches over time to learn the process. SVM, PLSR, Gaussian process regression (GPR) and regression trees (RT) are selected as potential model candidates. The focus is not necessarily to show which of this algorithm performs better, but more on the benefit of choosing a better suited approach for highly dynamic systems by learning from the real-time data and switching between the modelling algoithms. The proposed strategy to switch between different machine-learning methods in real-time based on their

performance, is similar to the switching strategy developed in (Tulsyan et al., 2014) for nonlinear state estimation and in (Tulsyan et at., 2018) for adaptive state estimation.

The proposed ML framework is tested using small-scale data retrospectively to forecast product concentration with a day advance. It is hypothesized that this kind of forecasting approach would give scientist to take a preventative action for potential undesired future trajectory.

2.METHODS

2.1 Statistical Modeling Approaches

Adaptive model selection algorithm is developed including, SVM, GPR, PLSR, regression trees (RT) and ensemble trees (ET).

Support Vector Machines. SVM received considerable attention due to its promising results in many different classification and regression applications. The SVM classifier developed by (Vapnik, 1995) tries to find the optimal hyperplane in n-dimensional space with the highest margin between classes.

Regression Trees. RTs are decision trees with binary splits for regression. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. (Loh, 2011).

Ensemble Trees are weighted combination of multiple regression trees.

Gaussian Process Regression. GPR is a "less parametric" supervised learning algorithm that rather than claiming a specific model, uses a Gaussian distribution to represent the data (Seeger, 2004).

Partial Least Squares Regression. Partial least squares is a multivariate regression method, especially convenient for large number of highly correlated data sets. The PLS models summarize the original data matrix (input variables X) to extract the most predictive information for the response variable (Y) and maximize the covariance between X and Y (Geladi and Kowalski, 1986).

2.2 Model structure and data

Continuous data received from pH, temperature (T) and oxygen sensor is combined with discrete daily data of mammalian cell culture. Discrete daily data includes nutrient and metabolite concentrations as well as cell cycle phase distributions obtained with offline flow cytometer. Continuous data is summarized into its mean, standard deviation and deviation from the set point to match daily discrete data. Product concentration (titer) is defined as response variable.

Expanded window is applied to the adaptive algorithm and model is re-trained whenever new performance data is available. Dynamic feature selection is performed using both variable importance projection (VIP) calculated by PLSR and predictor importance calculated by Ensemble Tress. The VIP values summarize the contribution of variables to the model and are calculated as a weighted sum of squares for each X variable by summing the squares of the PLS loading weights. Generally VIP values larger than 1 are accepted as important variables. When there's disagreement between two methods, union of the suggested features are used. All models are trained in real-time using combined data from four simultaneous bench scale cell culture experiments with different feed schedules and only predictions for two bioreactor runs are presented here (BR-1 and BR-2).

Each model is cross validated and optimized using Matlab machine learning and statistical toolbox. Hyperparameters of the models are reported in Appendix A. Model performances are compared and ranked using cross validated root mean squared error (RMSEcv).

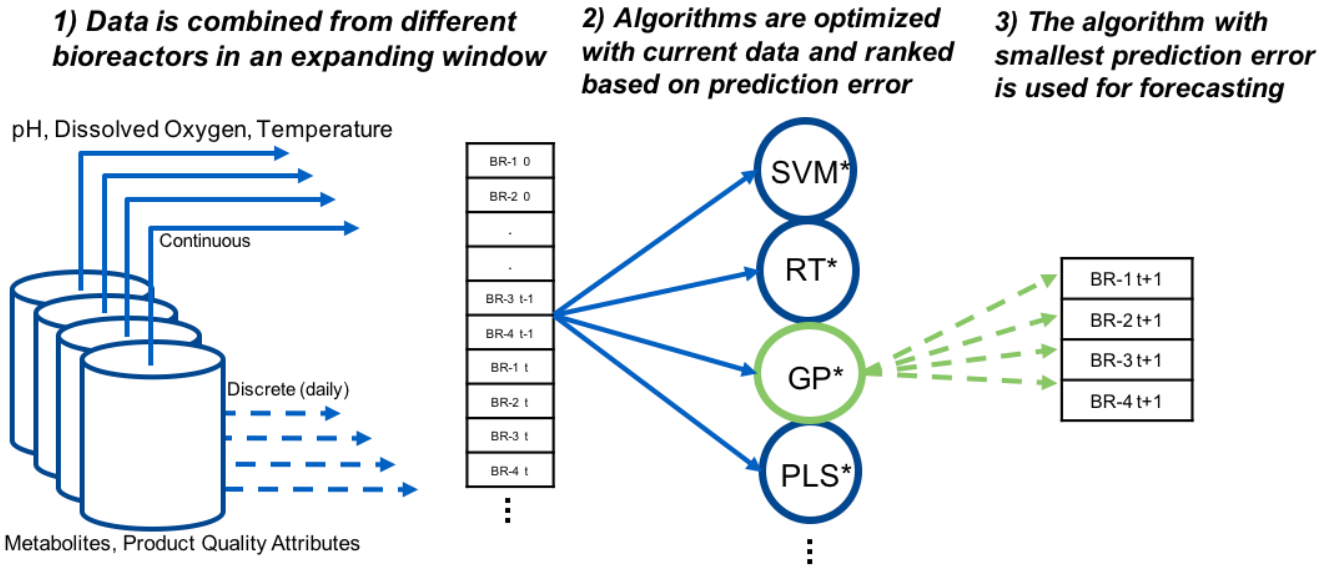

Figure 1 Schematic illustration of the real-time ML algorithm with expanding window

2.3 Experimental Study

All experiments were conducted in 3L single-use disposable CellReady bioreactors by Millipore with Amgen proprietary cell line and media. The process is a fed-batch process with working volume of the bioreactors control to less than 2L. The process was controlled in Sartorius BIOSTAT B-DCU QUAD controllers. Online data such as pH, DO and temperature from the bioreactor is captured in PI Historian. All analytical data were sampled either manually or via Nova Bioprofile Flex Autosampler. The cell culture samples were then analysed on the Nova Bioprofile Flex analyzer. The parameters include: Viable Cell Density (VCD), viability, pH, pCO2, pO2, glucose, glutamine, lactate, glutamate. ammonium, sodium, calcium and potassium. The experiment was performed in a fed-batch setting with pre-defined feeds added during the process. There are other scale-up steps associated with this experiment, but the data from these other unit operations were not used. Data from this experiment is from the production step in cell culture where protein production is made. Samples for protein concentration were taken daily and frozen in -20 ^{o}C freezer. The samples were then analysed in one batch post cell culture via high performance liquid chromatography (HPLC) to determine protein concentration. Millipore guava easyCyte 5HTTM flow cytometer with FlowJo version 10 (FLOWJO LLC) was used to fit the cell cycle populations.

3. RESULTS

An adaptive machine-learning framework is developed to capture the dynamics of a cell culture bioreactor and forecast product concentration one day in advance by combining data from different sensors retrospectively. The algorithm decides a modelling approach that is most representative of the current data.

The model performance is compared using dynamic feature selection vs. using the same subset of the variables at each time point for prediction. For dynamic feature selection VIP value above 0.5 and predictor importance value above zero are used (Fig. 2). Dynamic feature selection significantly reduced the prediction error (RMSEcv) for PLSR, GPR and SVM algorithms (Fig.3). However, there was no improvement for RT and ET methods. Dynamic feature selection algorithm is used for the rest of the study.

Fig. 4 illustrates the comparison of model predictions and experimental data for BR-1. As expected, prediction error starts high due to low number of training data and improves when the model collects more data to learn the process. GP ranked higher early in the process but later PLSR resulted with better performance. PLSR method required more data to learn and didn't develop a representative model until later in the process.

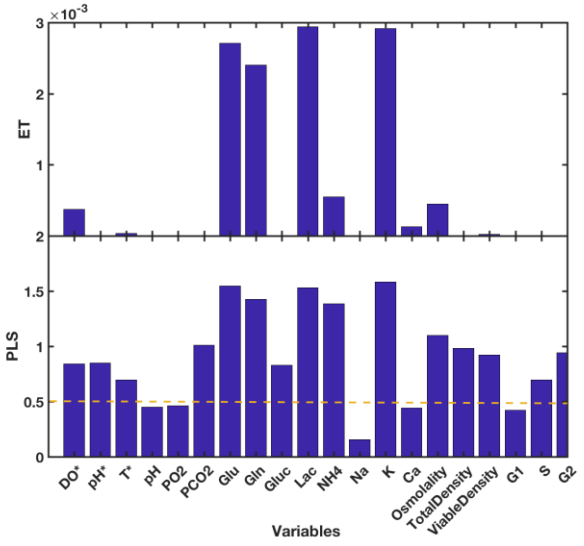

Figure 2 Variable importance plot suggested by ET and PLSR

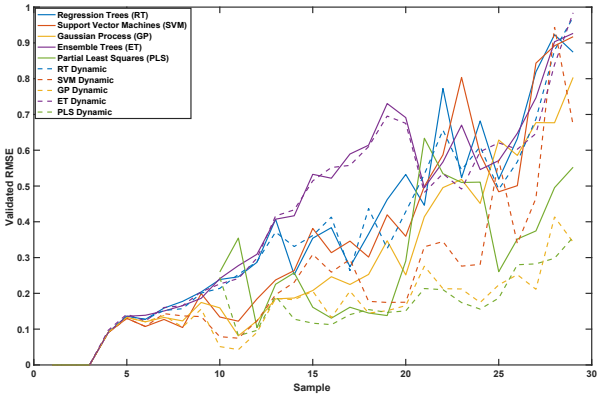

Figure 3 Comparison of RMSEcv for different algorithms using dynamic features selection (dashed lines) vs static features (solid line)

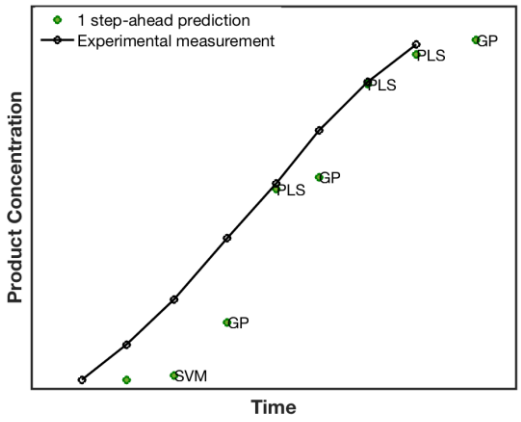

Figure 4 Comparison of model prediction versus experimental measurements for 1 day ahead prediction of product concentration for BR-2

BR-2 result showed similar trends as BR-1 and improved on prediction using PLSR later in the process (Fig. 5). GP was selected with SVM early on the process. Fig. 6 shows the average number that each variable was used during the simulation run for BR-1 and BR-2. Discritized mean T, DO, pH, daily Glutamic acid (Glu), glutamine (Gln), glucose (gluc), lactate (Lac), NH_4, K, viable cell density (VCD) along with the percentage of cells in G1 phase were used in each prediction to show the correlation of these variables with product concentration.

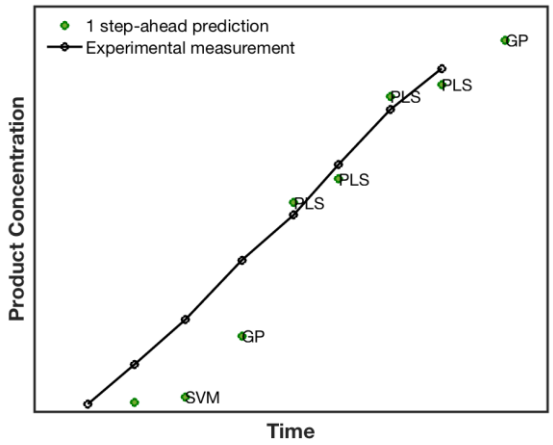

Figure 5 Comparison of model prediction versus experimental measurements for 1 day ahead prediction of product concentration for BR-2

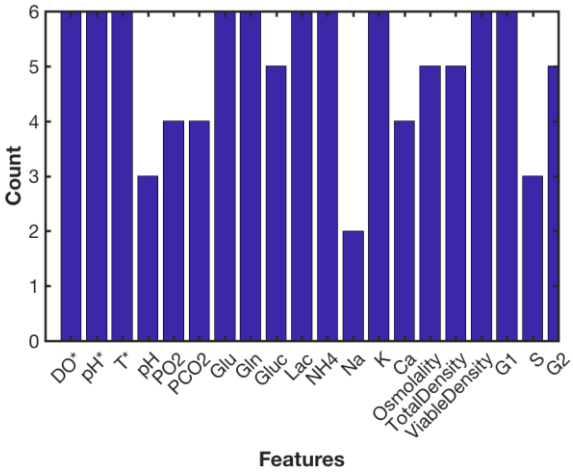

Figure 6 Average number of times the variables are used for prediction

Average RMSEcv is shown for all modeling algorithms in the framework (Fig 7).

In general, PLS and GP performed better as they resulted with the lowest error for each time step. Performance of RT and SVM was interchangeable while ET was most of the time gave the highest error.

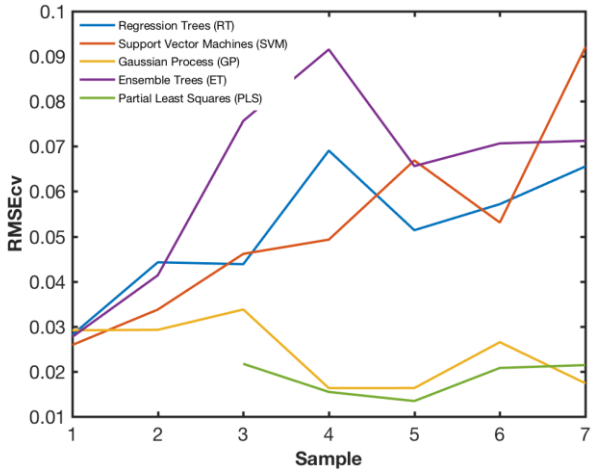

Figure 7 Compariosn of cross validated RMSE values for different algorithms

4. DISCUSSION

In biopharmaceutical industry, there is a need for intelligent systems that learn in place, in real-time from the history available, from the current dynamics of the process and inform us about the future. In this study, we developed such framework to capture the dynamics of the system by switching between different ML algorithms.

Limited number of ML algorithms are used as the focus was on the real-time learning and ranking of the algorithms. One can easily plug-in other ML approaches that might be more predictive for a specific domain.

The algorithm is tested using two small-scale bioreactor experiment including only 8 sample points using an expanding window. An alternative would be using a moving window to include recent data and forget the data early in the process that might not be relevant anymore. For longer processes such as continuous manufacturing moving window can be a better approach to be implemented.

Daily cell culture data especially for measurements such as viable cell density can be quite noisy. In this work no filtering algorithm is applied as the raw data directly fed to the models. When working with small data sets noise can be detrimental to the models, there are filtering approaches that could be used to reduce the adverse effect of noise (Vaseghi, 2001).

Machine learning algorithms work best dealing with larger data and extracting the most useful information. We have assumed there was no historical data available for this process at the beginning of the run. If such data become available prediction horizon can be updated to predict further future or final product quality.

During this work, we used the cell culture process data from different sources offline including cell cycle distribution as it is a critical piece of knowledge to evaluate the health of the culture. Automated flow cytometer (Abu-Absi et al., 2003) or

in-silico approaches can be used to provide this information in real-time (Bayrak et al., 2016). Similarly, product concentration or product quality attributes are available near real-time with micro sequential injection technologies (Wu and Wee, 2015).

Upon implementation of such a ML framework, cell culture scientists could forecast product quality attributes and current correlations in real-time. Such a framework could be deployed to any process with minimal effort and can reduce the number of wet lab experiments significantly. Commercial transfer of this methodology would also help to build the bridge between small-scale and manufacturing data.

5. CONCLUSIONS

We have demonstrated that real-time machine learning algorithms can provide insights to the future of the current run in mammalian cell culture bioreactors. Model predictions can provide to improvement in monitoring, control and optimization strategies for cell culture process thereby leads to more robust process development and manufacturing of important biologics.

REFERENCES

Absi, N. R., Zamamiri, A., Kacmar, J., Balogh, S. J. & Srienc, F. 2003. Automated flow cytometry for acquisition of time-dependent population data. *Cytometry A,* 51**,** 87-96.

Agarwal, H., Rathore, A. S., Hadpe, S. R. & Alva, S. J. 2016. Artificial neural network (ANN)-based prediction of depth filter loading capacity for filter sizing. *Biotechnology Progress,* 32**,** 1436-1443.

Bayrak, E. S., Wang, T., Jerums, M., Coufal, M., Goudar, C., Cinar, A. & Undey, C. 2016. In Silico Cell Cycle Predictor for Mammalian Cell Culture Bioreactor Using Agent-Based Modeling Approach. *IFAC-PapersOnLine,* 49**,** 200-205.

Geladi, P. & Kowalski, B. R. 1986. Partial least-squares regression: a tutorial. *Analytica Chimica Acta,* 185**,** 1-17.

Jia, L. & Sun, Y. 2017. Protein asparagine deamidation prediction based on structures with machine learning methods. *PLOS ONE,* 12**,** e0181347.

Le, H., Kabbur, S., Pollastrini, L., Sun, Z., Mills, K., Johnson, K., Karypis, G. & Hu, W. S. 2012. Multivariate analysis of cell culture bioprocess data--lactate consumption as process indicator. *J Biotechnol,* 162**,** 210-23.

Loh, W.-Y. 2011. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 1**,** 14-23.

Schmidberger, T., Posch, C., Sasse, A., Gülch, C. & Huber, R. 2015. Progress toward forecasting product quality and quantity of mammalian cell culture processes by performance-based modeling. *Biotechnology Progress,* 31**,** 1119-1127.

Seeger, M. 2004. Gaussian Processes For Machine Learning. *International Journal of Neural Systems,* 14**,** 69-106.

Stefik, M. J. 1985. Machine learning: An artificial intelligence approach: R.S. Michalski, J.G. Carbonell and T.M. Mitchell, (Tioga, Palo Alto, CA); 572 pages, $39.50. *Artificial Intelligence,* 25**,** 236-238.

Tulsyan, A., Huang, B., Gopaluni, R. B., & Forbes, J. F. (2014). Performance assessment, diagnosis, and optimal selection of non-linear state filters. *Journal of Process Control*, *24*(2), 460-478.

Tulsyan, A., Khare, S., Huang, B., Gopaluni, B., & Forbes, F. (2018). A switching strategy for adaptive state estimation. *Signal Processing*, *143*, 371-380.

Vapnik, V. N. 1995. *The nature of statistical learning theory*, Springer-Verlag New York, Inc.

Vaseghi, S. V. 2001. Adaptive Filters. *Advanced Digital Signal Processing and Noise Reduction*. John Wiley & Sons, Ltd.

Wu, C. H. & Wee, S. 2015. Micro sequential injection system as the interfacing device for process analytical applications. *Biotechnol Prog,* 31**,** 607-13.

Appendix A.

Table A.1 Hyperparameters of modelling algorithms

Algorithm	Hyperparameters
ET	Method= 'Bag' Number of Learning cycles= 30 Minimum Leaf Size= 8
SVM	Box Constraint= 0.24 Epsilon= 0.024 Kernel Scale= 4.4
GP	Kernel Function= 'SquaredExponential' Beta= 0.3168 Sigma= 0.0079
RT	MinLeafSize= 4
PLS	Number of components=1 0