November 21, 2018

MULTI-FACTOR ADAPTIVE MACHINE LEARNING EVALUATION IN PRODUCTION ENVIRONMENTS

HP INC

Multi-factor Adaptive Machine Learning Evaluation in Production Environments

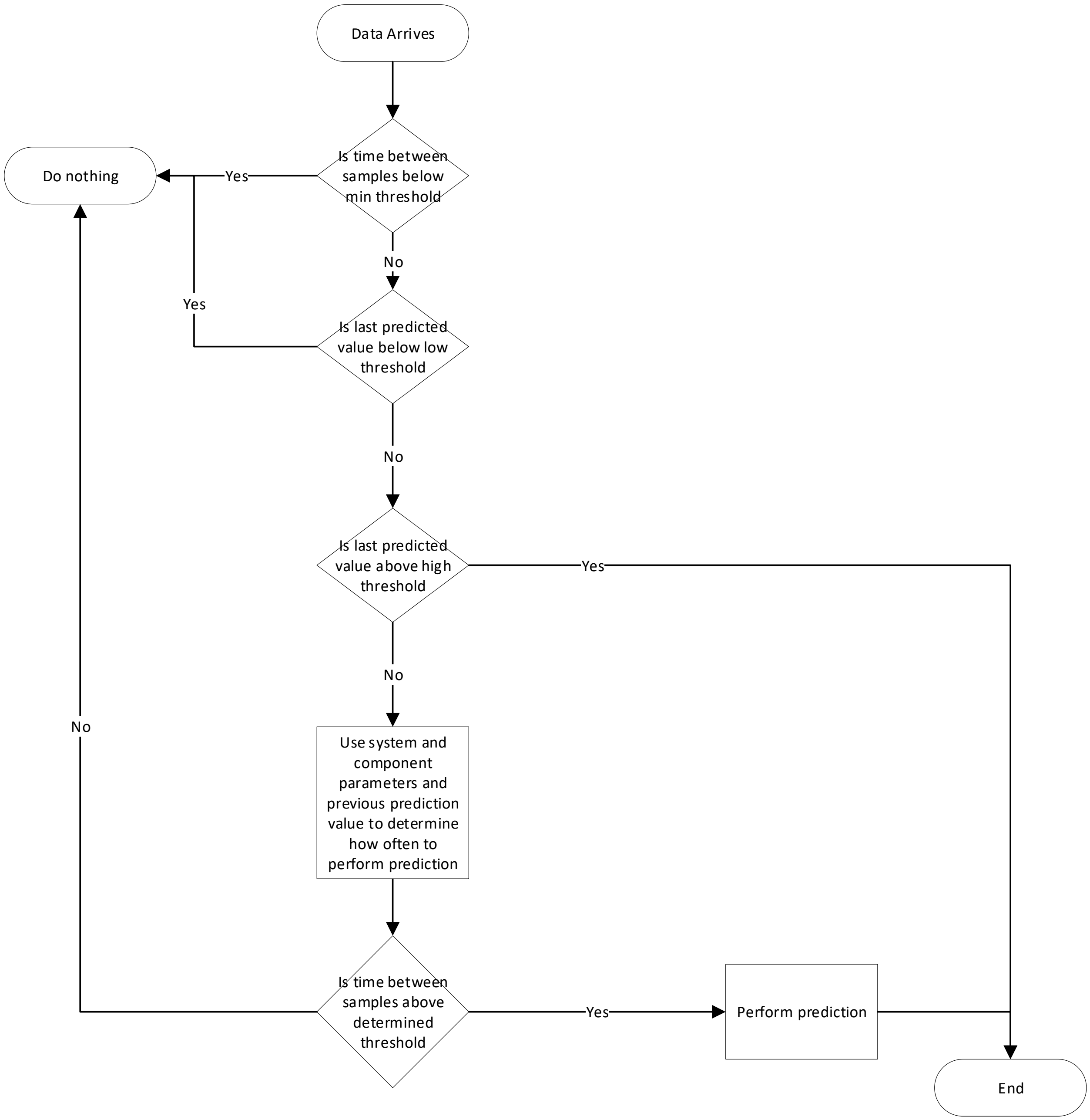
In machine learning applications, a system takes a data set and runs it through the machine learning model to create a prediction. It is necessary for some systems to do this on a continual basis for example, a computer vision system continually evaluates the environment to recognize objects. In other systems it is not necessary to run data through the model every time new data is available and may not result in a significantly different prediction if it was. In these systems, it would be more cost-effective to space out runs of the model. Since running a machine learning model can be expensive, having a method to reduce the frequency of runs can provide significant savings to a business. In a continuously monitored system, every time data arrives a new prediction is done and in many cases the prediction is not significantly different than the previous prediction. This consumes resources which affects the responsiveness of the system and increases operation cost. By only doing predictions when it's likely the new prediction will be significantly different, the system improves responsiveness and reduces the cost of operation. In a batch system, predictions are done on a fixed schedule. This method would change the system such that it would only add a record to the batch when it's likely the new prediction will be significantly different. This also would improve system responsiveness and reduces cost of operation.

In a system using machine learning, when the data arrives as signal is sent to another process indicating data is ready. This process will then check if a prediction should be performed. Items that could be checked are: last time a prediction was done, the value of the last prediction, system data, and data about the item being predicted. The time since last prediction and the value of the last prediction can be used in all cases. If the last value for a prediction is above the high threshold then run the machine learning model at the minimum time spread value (this could be as frequent as each time data arrives). If the last predicted value is below the low threshold then run the machine learning model at the maximum spread time value. If the predicted value is between the high and low thresholds, then use the system parameters, component parameters and the last predicted value to determine the interval at which the prediction should be done. Using that interval determine if the time between predictions has exceed the interval and if so then perform the prediction (see diagram below). Some examples:

Example 1. The prediction is the probability of failure of a fuser on printing device (this is a long-life component). If the fuser has a percent life remaining near 100% then predict no more frequently than once every two weeks. As the percent life remaining approaches 50% then predict more frequently. The relationship between percent life remaining and prediction frequency can be either a linear or nonlinear relationship.

Example 2. The prediction is the probability of failure of a hard disk drive. If the previous probability of failure is near 0% then only run the prediction model weekly. As the probability of failure approaches 100% then predict more frequently until the predictions occur on every time the data arrives.

Example 3. The prediction is the probability of failure of a scanning device on a digital scanner. If the number of pages that have gone through the scanner since the last prediction is below a threshold (i.e. 10000 pages) and the previous probability of failure is less than 50% then do not predict with this data collection. If the number of pages is above the threshold or the probability of failure is greater than 50% then predict on this data collection.



Disclosed by Daniel Siddall, Darrel D Cherry and Anton Wiranata, HP Inc.