Article can be accessed online at http://www.publishingindia.com

Analytically Yours:

Analysis of Data Streams

Arnab Kumar Laha*

One of the four Vs of Big data is 'velocity' which refers to the fact that in many applications, data is not static but continuously flows into the system (often at a very high rate). Such continuously flowing data is termed as Streaming data and is generated by various sources such as surveillance cameras, sensors in machines such as aircraft engines, tractors, vehicles and mobile phones, atmospheric systems, mass production systems, transactions such as those of a credit card system etc. In some applications such as credit card fraud detection, intrusion detection or preventive maintenance it is important that we are able to analyse the data stream in real time. There are two major challenges associated with analysis of data streams which are not present when dealing with static data. Firstly, it is not possible to work with the 'whole data' as the data keeps flowing into the system and secondly, the nature of the data changes over time- a phenomenon often referred to as 'concept drift'. Thus, analysis of streaming data requires techniques different from those used for static data.

Since the data flows continuously and often at a very high rate it necessitates the use of techniques that allow us to update the results quickly as new data is accumulated. Specifically, techniques that require use of the entire data available at each point of time or that require multiple passes over the entire data set are often not suitable for use with streaming data. Moreover, as concept drift is commonly present in streaming data we need to monitor the results (or output) quite closely to detect the occurrence of a change in the data generating system so that the model used can be updated. When concept drift is present in the streaming data, it is not even appropriate to use the entire historical data for building the model. Instead, researchers have suggested several alternative methods. Among these, the use of a 'data window' is the most popular. In this approach the model is built using a subset of data typically the most recent. Once a model is built, its predictive performance is tracked and when the model's performance deteriorates, it is re-built using the most recent window of data. The window size i.e. the number of recent observations to be included in the window, is chosen keeping in mind both the accuracy of the predictions as well as the computation time required to build the model. The second factor is important for real time applications.

Let us illustrate the above ideas using a simple example. Suppose we have streaming data about two related variables X (say, distance travelled by a passenger in a taxi) and Y (say, taxi fare) and we are interested in predicting Y based on X. We consider a simple linear regression model and compare two strategies: (A) build the model with the first 100 observations and use the same for predicting, and (B) build the model with the first 100 observations and then keep rebuilding the model with the latest 100 observations whenever 500 predictions are completed. We compare these strategies in two different scenarios: In scenario I, there is no concept drift while in scenario II, concept drift is present. Specifically in scenario I, the data generating mechanism $Y_i = 1 + 2X_i + \in_i$ is where $\in_i \sim N(\mu = 0, \sigma = 2)$, i = 1,…,3000 whereas in scenario II, the data generating mechanism is

$Y_i = 1 + 2X_i + \in_i$ where $\in_i \sim N(\mu = 0, \sigma = 2)$, $i = 1,\ldots,1000$,

$Y_i = 1 + 1.5X_i + \in_i$ where $\in_i \sim N(\mu = 0, \sigma = 2)$, $i = 1001,\ldots,2000$ and,

$Y_i = 1 + X_i + \in_i$ where $\in_i \sim N(\mu = 0, \sigma = 2)$, $i = 2001,\ldots,3000$.

We begin our discussion by comparing the performance of two strategies in scenario I. We use the Root Mean Square Prediction Error (RMSPE) as the measure of performance. It is computed as

* Indian Institute of Management, Ahmedabad, Gujarat, India. Email: arnab@iima.ac.in

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where $\hat{y}_i$ and y_i are the predicted and actual value. We find the RMSPE values of the two strategies are very close to one another as would be expected in this case ($RMSPE_A$ = 1.98 and $RMSPE_B$ = 1.99). Thus, periodic rebuilding of the model does not yield any benefit in this case.

Now, let us examine how these two strategies perform in scenario II. We find that the RMSPE value of strategy A is far more than that of strategy B ($RMSPE_A$ = 6.95 and $MSPE_B$ = 2.35) indicating that strategy A performs quite poorly compared to strategy B. Thus, we find that the periodic rebuilding of the model used in strategy B had a positive impact in terms of improving the prediction quality. In general, periodic rebuilding of the model is a recommended strategy when dealing with streaming data where concept drift is suspected to be present.

A natural question that arises at this stage is how one knows that a rebuilding of the model is required. One way is to use the prediction errors. Intuitively we feel that if the prediction errors become larger than expected, then we should rebuild the model with current data. A widely used tool for handling such problems is the control chart introduced by W. Shewhart in the context of industrial quality control. We now discuss how a control chart can be used for monitoring the model output. For the purpose of this illustration, we restrict ourselves to scenario II. As before we build the model using the first 100 observations and obtain the residuals (residual = actual value - fitted value). These residuals are used for creating an individuals control chart for monitoring prediction errors. Since the average value of the residuals in linear regression is 0, we can use that as the centre line of the individuals control chart. The Upper Control Limit (UCL) in the individuals control chart is set to 3s and the Lower Control Limit (LCL) is set to –3s where s is the standard deviation of the residuals. This individuals control chart can be used for monitoring the model performance as follows. Each prediction error is plotted on the control chart taking care that the same sequence as that of the arrivals of the observations is maintained. If a prediction error falls above the UCL or below the LCL we decide to rebuild the model. Fig. 1 gives the prediction errors plotted on the control chart which is created using the R package qcc. As can be seen from the chart that the prediction error of the rightmost observation is above the UCL indicating the need for rebuilding the model.

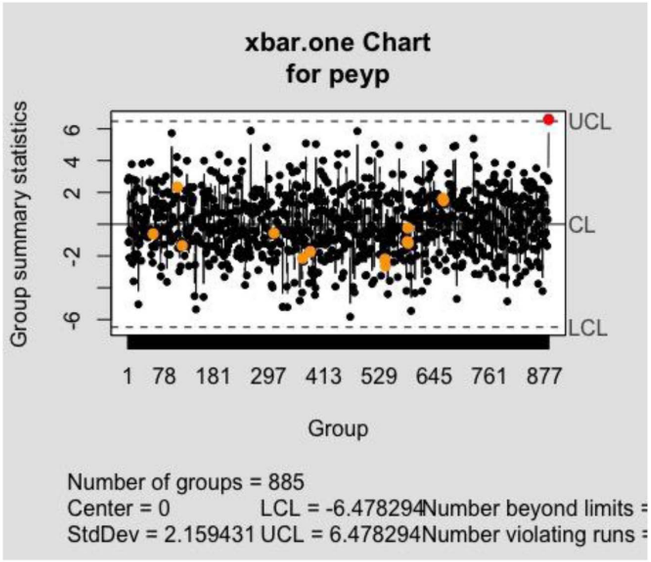

Fig. 1: Prediction Errors Plotted on the Control Chart

In summary, we can say that when dealing with streaming data with possibility of concept drift it is important for us to track the performance of the model and rebuild it as and when necessary. The individuals control chart for prediction errors can provide guidance regarding whether the current model is adequate or there is a need to rebuild the model with current data.