**Chi-Square Test**

We tried Chi-Square test to detect changes in binary and categorical datasets. Chi-square compares distribution of two windows and test of the difference is significant. We tried varying the window size from 10 to 50 and varying the overlapping between the window size from 0 to 40. The file 'ChiSq/results.txt' shows that only window size = 50 and overlap = 40 does not give a false alarm on negBinary.txt and negTriple.txt.

As a result, we keep this Chi Square test in the binary and triple model as the second line of defense after the Shewhart. We set the parameters to be quite strict (i.e. low significant level, high overlapping) in order to minimize false positive.

**Kolmogorov-Smirnov (KS) Test**

Kolmogorov-Smirnov test is another test that we use to check if the concept changes significantly. We tried KS test with our continuous data (i.e. mean and variance change). We chose KS test over Student's t-test because KS test is a non-parametric test and is not limited to data with Gaussian normal distribution.

From 'KS/results.txt', you can see that we tried varying window size and the overlapping as we did with Chi Square. However, when the window is 20 or 30, KS test did a good job in datasets that don't have change. We get the best detection result at the window size of 20 and no overlapping. It detected all changes except posShiftUpVar\_60.txt. Although KS detection is quite late (i.e. 97 vs. 90 in posShiftDownMean\_90), we think late is better than never.

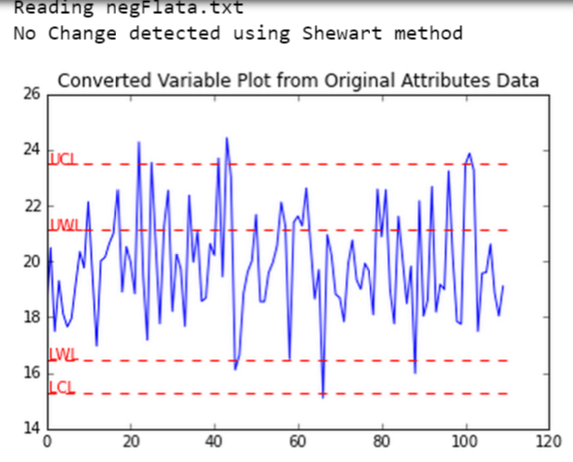
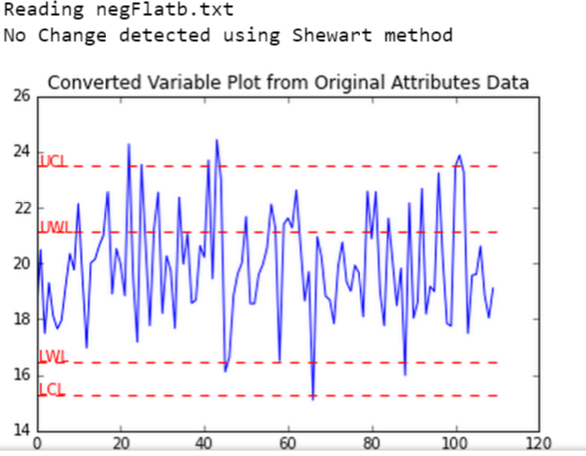
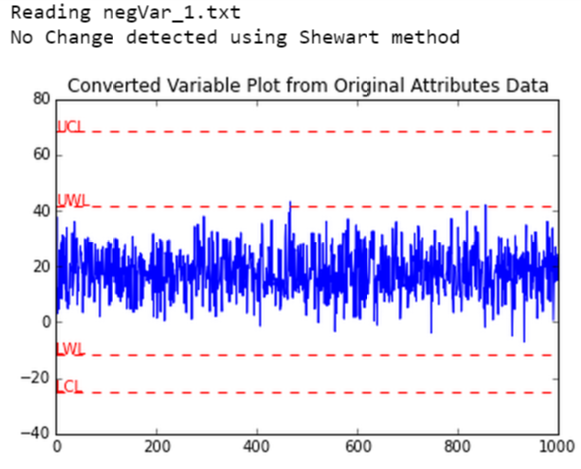
As a result, we chose to use KS with a window size = 20 as our third line of defense after Shewhart and CUSUM. This should give us another safety net in case Shewhart, which is quite precise in change position, fails to detect change.

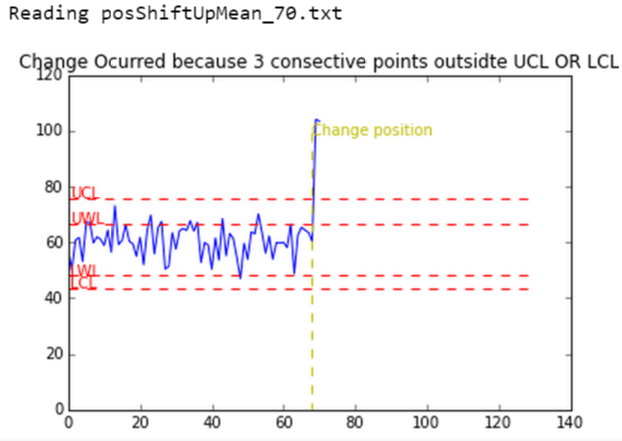
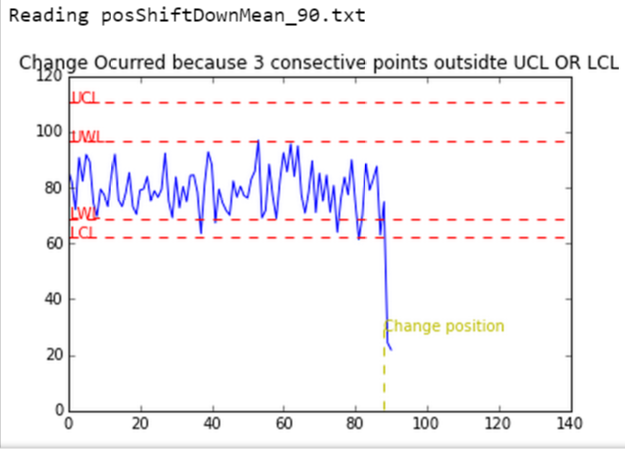
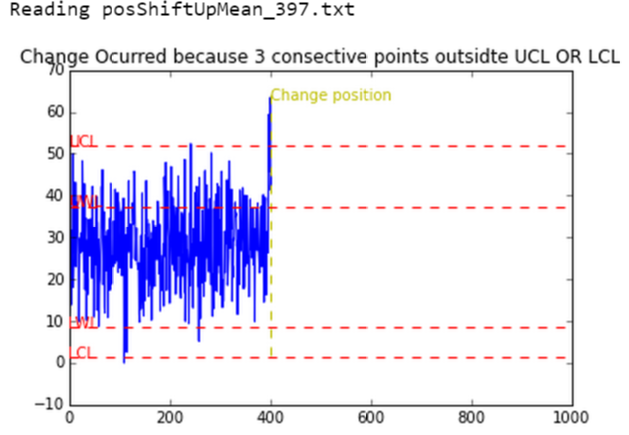
**Shewhart Control Chart**

Shewhart charts is used to determine if a process is in a state of statistical control, and if it first was and then not, that signals a concept change. The method we implemented choose the first 5 samples to set up the baselines: calculating the mean, upper control limit UCL ( mean + 4\*std), upper warning limit UWL( mean + 2\*std), lower control limit LCL ( mean - 4\*std) and lower warning limit LWL( mean - 2\*std), the limits relationship to mean value in terms of number of standard deviation are selected using training data and tuned with test data.

Once baselines are set and stored, this method just compares the next 5 points to the baseline limits, if there are 3 consecutive outside UCL or LCL, a concept change is signaled, and the position starting to change is 3 position past, if there are 5 consecutive points outside UWL or LWL, ,a concept change is signaled, and the position starting to change is 5 position past. The advantages of Shewhart Control Chart is the relatively small windows it uses: 10 ( the first 5 points to calculate baselines and store baseline limits, and other 5 for continue comparing against baseline limits), also it is simple and easy to implement.

Shewhart Control Chart can detect most of provide sample data and generated sample concept change exists or not correctly, and we find out it is most effective when there is a relatively significant phase change, and thus we choose to use this method as the first detecting method and here are a few of the examples we tested below:

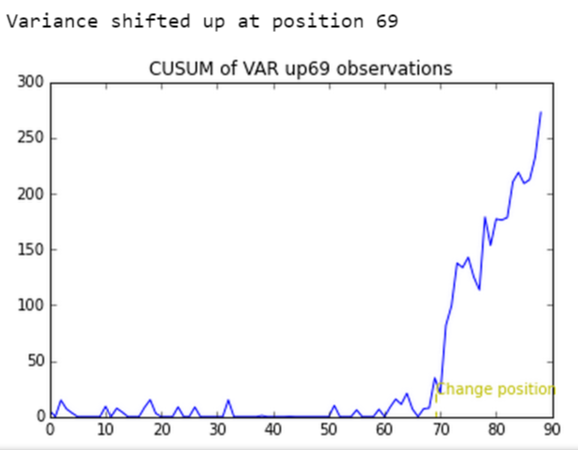
 



**CUSUM Method**

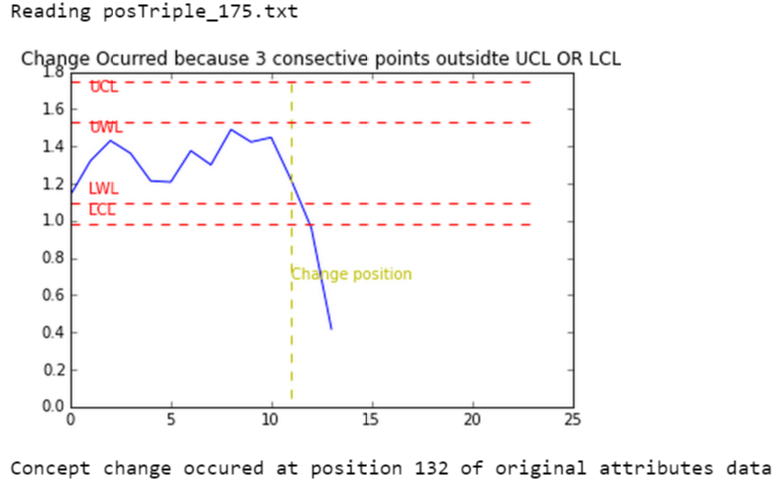
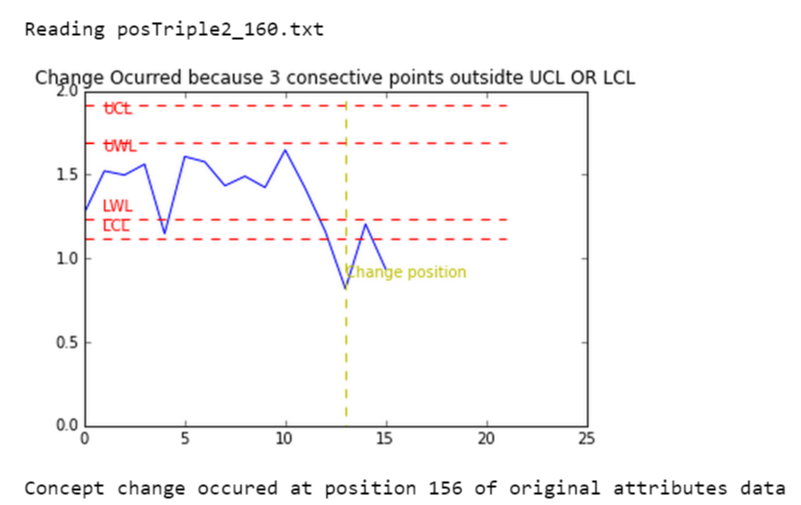
A CUSUM chart is a time-weighted control chart that displays the cumulative sums (CUSUMs) of the deviations of each sample value from the target value. Because it is cumulative, even minor drifting in the process mean or variance will lead to steadily increasing or decreasing cumulative deviation values. We used [Hawkins CUSUM Method](http://rmgsc.cr.usgs.gov/outgoing/threshold_articles/Hawkins_Zamba2005b.pdf) and in Douglas C. Montgomery Statistical Quality Control 6th Edition Chapter Time-Weighted Control Charts. There are following tunable Parameters: k : Allowance value, a threshold that signals a possible shift starts at the current position; H : If CUSUM value goes beyond this value, conclude a shift has occurred and the shift starts at the last point whose CUSUM is larger than k Allowance value; Separate k and H values are used for mean and var shift detections.

Since CUSUM works best in detecting small drifting in process, we used CUSUM as the second detection method to scan data that failed to be detected any change by Shewhart method.

Example of Variance change detected at position 69 in posShiftUpVar\_70.txt

**Attribute Data Handling**

Besides Chi-Square method we use to scan attribute data like binary and triple as the 2nd scanning method, the first detection method we used is to convert attribute data to continuous data, and then apply Shewhart, CUSUM, K-S method to scan any change concept occurred in the continuous data converted from original attribute data. We assign random uniformly distributed values to a, b, c at different range: a: range 0 to 1, b: range 1 to 2, c: range 2 to 3. This conversion allows us to use the continuous data detection methods we already implemented and it works best when there is a significant occurrence frequency for some of the attribute data occurred. As following examples shown, we converted posTriple\_175.txt, PosTriple2\_160.txt ‘s attribute data to continuous and then used Shewhart method detect position change at 132 and 160.

**Other Methods**

We also tried some other methods. One is Student's t-test. However, we are not sure whether the actual datasets are normally distributed or not and the small window size is usually not suitable for t-test. We decided not to implement t-test as part of our main model. Another method that we considered but later did not include in the model is the sequential probability ratio test (SPRT) which is the predecessor of CUSUM. Another method we tried is simple sliding window comparison, we use two consecutive window with size of 15 each, if the 2nd window’s mean or variance is significantly different than the first window, for example we picked 30% mean and var shift as threshold, then a change is detected and alert, this method gives false alarms to some files.

**Integrated Strategy**

Above all the methods we tried, each of them has its own advantages and disadvantages, and to combine them in a deliberately designed way can help us detect changes at different situations, avoid pitfalls and simply become a more flexible, powerful integral method. Based on this consideration, the final integral approach we take are as following:

For continuous data: first detection method is Shewhart which works best for large drifting, then CUSUM if Shewhart fails to detect change, and then Kolmogorov-Smirnov method which is the most sensitive method for minor concept shift.

For attribute data: first we convert them into continuous data using method described in Attribute Data Handling paragraph, then we apply Shewhart and Chi-Square method consecutively to scan the converted data for any concept change. The integral approach we tested and tuned were able to give us most accurate label for provided and generated sample data.

**Test Data**

After tested the model with the datasets provided, we generated our own datasets to test our model. We use normal distribution to randomly generate data for mean and variance changes. Our datasets have 1000 observations. We know that our variance and mean changes can be difficult to detect because means and variances are randomly selected. Our attribute data, however, follow a uniform and Diriswchlet distribution.

The result can be seen at Dingchao\_and\_Pipat.txt file.