Dynamic Big Data Drift Visualization of CPU and Memory Resource Usage in Cloud Computing

Tajwar Mehmood, Seemab Latif

Dynamic Big Data Drift Visualization of CPU and Memory Resource Usage in Cloud Computing

Tajwar Mehmood and Seemab Latif

School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST), Islamabad, Pakistan.
(tmehmood.phdcs17seecs, seemab.latif)@seecs.edu.pk

Abstract. Drift Visualization gives better insight into the nature of changes in the data distribution. Cloud trace is dynamic and generated at a very high pace, it's a serious problem that needs more deep discussions. Drift in cloud resource usage can cause low resource efficiency. In order to achieve optimal resource utilization, cloud provider needs a prediction model based on the data insights. Efficient drift detection optimizes the model prediction. Changes in data can cause these models to reduce their accuracy over a period of time. The focus of this research is to visualize the drift in the cloud at the cluster level. These visualizations will help cloud providers in understanding major factors contributing to the drift. In this paper, Cluster-based visualization using k-means is used to show the drift in the cloud.

Keywords: Concept Drift · Drift Detection· CPU Usage · Memory Usage · Drift Visualization

1 Introduction

Cloud resource utilization is an open research area that needs more attention. Multiple types of users are sharing over pay as you go model, each with its own unique demands [8]. Each workload consumes varying resource usage. Cloud resource usage traces are rapidly generated as a stream [19] along with complex changes in the distribution [11]. Cloud provider promises to deliver services to its user with the flexibility of scaling both up and down [6]. User demand can change due to sudden, seasonal, or recurring trends. In order to maintain its reputation and promise in a constantly changing environment, clouds need an adaptive prediction model. A good adaptive model can only be designed if the researcher has deeply analyzed the drift in the respected domain. Visualization provides an easy and effective way to comprehend the drift occurrence.
Static models are designed on the basis of the assumption that no changes are happening in the distribution. They do not have the capability to deal with changes. They are learning from the limited amount of training data. Testing data sets are prone to changes due to many factors. The training set can miss a

lot of information about the real scenarios. In such a dynamic environment, we need more than a simple prediction model. Whereas, an adaptive model with constant updating is time-consuming and loss of previous information. Frequent model updates can be prevented by using a drift detector but still do not provide information about the type of changes.

Data distribution changes also referred to as concept drift can fail a predictive model. A change in the distribution of input values can cause a change in the labels mapping. An attribute dependency on label class could have been changed or change can happen within the individual attribute. These changes can occur at a different pace of time. A drift that occurs over a shorter period is a sudden drift. Whereas, a drift that occurs over a longer period of time is the gradual drift. Further, drift can also be categorized as virtual or real. A change in the class label is virtual drift, changes in parameter distribution are referred to as real drift. Concept drift is different from anomalies which can be seen as outliers. Anomalies are caused due to critical data leak or deviating from the normal behavior [3]. Whereas, drift is the changes in distribution that can affect the accuracy of the present prediction model but are still a normal behavior.

Different drift detection techniques can detect the drift type but provide no insight into the pattern. In order to understand patterns from these fluctuating usage traces, we need to understand the reasons that are causing these changes. Visualization is one of the ways to understand the patterns more deeply but also can help in identifying the reasons for the drift. Visualization is an easy and more convenient way to depict the changing pattern that is causing different types of drift. Data Charts and maps can be drawn using different techniques to extract unique information. Statistical analysis can also extract patterns from the data and get detailed insight into the distribution. We needed publicly available resource usage data to continue with this research. There are some cloud trace data publicly available that gives a real insight into the cloud providers' usage traces.

The major contribution of this paper is the visualization of sudden drift in the cloud. This visualization is achieved in three steps. First, the CPU and Memory usage are examined individually. In the next steps, the correlation of both resources that are causing changes in usage is identified. This is further shown using a cluster-based visualization. Lastly, identification of external factors that can indirectly affect the sudden drift in the distribution. This analysis can help researchers design better drift detection techniques for cloud providers.

The paper is organized into three sections. First is the literature review related to the sudden drift in the cloud. In the second section, problem analysis is covered along with the publicly available cloud usage traces. The third section shows the deep analysis of the sudden drift visualization. In the end, the paper is concluded with the future work.

2 Literature Review

Most research in the cloud is done on resource usage prediction rather than searching for the reason for usage changes. Load balancing and other mechanisms are already being used by cloud providers to deal with efficient resource utilization but drift detection is ignored. A usage trace can undergo any type of change in the distribution with respect to time, thus, affecting the prediction. A sudden drift exists nearly in all types of domains. There are many internal and external factors contributing to the causes of the drift. In each domain, different factors are causing these changes in the distribution.
Visualization is an easy way to detect and find the reasons for the changes in the concept. Visualization helps a complicated concept to understand in an easy way. Understanding, analyzing, and identifying patterns is one of the major advantages of visualization. These visualizations can vary from simple to complex techniques. Visualization techniques can be divided into three major categories on the basis of the features i.e. Uni, Bi, and Multi variants. Uni-variant is an analysis of a single feature whereas Bi and Multi Variants are combinations of more than one different feature. A line chart variation is being utilized by [16] and [9] to visualize drift. A simple line plot along with the contribution of each attribute is added. An individual as well as aggregated positive and negative mean is represented above the line plot using dashed lines depicting the drift[9]. The importance of visualizations to study drift using scatter plots is focused on by researchers in the medical informatics domain [13]. Brush's parallel histogram is used by Kelvin et al. [10] to visualize the concept drift. It can represent multi-dimensions of data using parallel coordinates. It can transform multidimensional data relationships in two dimensional understandable form [4]. Marlon et al. [7] Anton et al. [9] both have done work in the business processes. Initially, [7], worked on sudden and gradual drift visualization using the adaptive window techniques. Later, [9] proposed a visual drift detection mechanism based on drift maps, charts, and graphs. Drift visualization in time series data needs more detail than a single visualization technique. Wang et al. [15] proposed a complete system to visualize drift at individual feature levels as well as the effect of the feature on other features and labels. They combined the effect from three different data sources depicted using heat maps.

3 Problem Analysis

Drift Magnitude is easy to visualize but it only shows the drift at each instance level. Group level visualization helps to show the combined effect of drift. Usage clusters can help to see data in terms of high, medium, and low resource usage and can be referred to as capacity groups. Creating capacity groups allow us to compare two different cloud datasets at the same level. A change in these capacity groups can lead to the detection of drift.
Clustering is a technique that easily divides continuous data into categorical forms. This categorical conversion helps in the visualization of the continuous

data in a more understandable way. We have used K-means clustering to find the capacity groups. K-means is the most suitable clustering technique for this visualization. K-means can process large amounts of data efficiently [14]. K-means allows forming of clusters on the basis of Euclidean distance and improves its performance using the Elbow method. Different numbers of clusters are used based on the Elbow method to find the optimum number of clusters [2]. A heuristic is used to select the right value of K. Distortion is the distance from the center of clusters and Inertia is the distance of all points from the center of each cluster. On the basis of distortion and inertia, the optimum value of K is selected. In this paper, each cluster in the chart is represented using a different color.

3.1 Datasets

Two cloud cluster datasets are used to compare utilization and drift patterns. This visualization is performed offline and can be done in real-time as well. Google Cluster Usage [12] and Alibaba [18] Trace are used in this study. Both dataset sets are widely used by the cloud research community. Google Usage Trace can be considered as a benchmark dataset for cloud usage trace understanding. Google usage trace is freely available for research purposes. This is a dataset of a one-month duration. It contains resource usage information of a single cluster. Three main types of resources are monitored i.e. CPU, Memory, and Disk I/O. The values of each resource are normalized from 0 to 1. Zero means low usage and it increases high while moving towards the 1. Alibaba Cluster Usage Trace also released a cluster usage trace to help researchers and students. Alibaba is also the largest Cloud computing platform available. They are also providing data related to CPU and Memory utilization at a task level. All resource usage information is normalized from 0 to 1 in order to compare values with the google usage trace. We have selected a limited number of instances in both datasets to visualize the drift. The current window size of 50000 instances is selected which can be varied to see drift over a short or long time ranges. Both datasets are from different time periods but both follow the non-Gaussian distribution and suffer from drift in distribution. Figure 1 and 2 shows the complete variation in the datasets. Google Data was originally normalized from a 0-1 scale but Alibaba was not normalized. Normalization is necessary in order to compare the two dataset's highest and lowest values at the same scale using the min-max method. Datasets characteristics are summarized in the Table 1.

4 Visualization and Analysis

Alibaba and Google Datasets are compared with respect to CPU and memory resources. A CPU usage of a cloud cluster in Figure 1 visualizes CPU consumption variance of tasks with respect to time in which we can see prominent spikes in CPU resource usage. CPU usage traces of both datasets have sudden spikes

Table 1. Dataset Characteristics

Datasets	**Alibaba**	**Google**
Selected Number of Instance	10,48575	10,48575
Normalization range	0 to 1	0 to 1
Duration	2 Months,1 Cluster	1 Month,1 Cluster

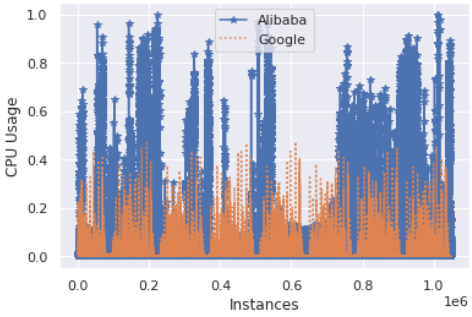

Fig. 1. CPU Usage Variation

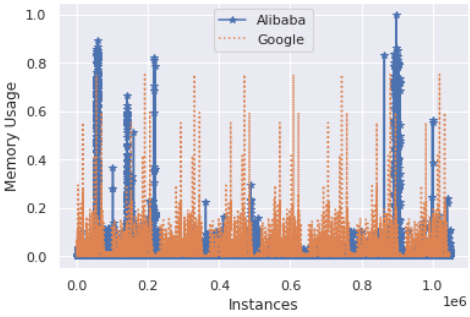

Fig. 2. Memory Usage Variation

after a certain time. Along with that, there are certain periodic drifts in Google usage data. A cyclic pattern of sudden peak and low usage can is very prominent in the Google CPU usage trace. Whereas, the pattern of Alibaba's CPU usage is very unpredictable. Alibaba resource has higher usage as compared to Google. Google memory usage is given in Figure 2, a similar recursive pattern to CPU can be visualized because both CPU and Memory usage has a very strong correlation, mentioned at the end of the current section. Sudden decline and peaks of google CPU and Memory are also coordinated in Figure 1 and 2. Unusual spikes can be seen in Alibaba whereas a recursive pattern can be visualized in Google usage variations. In Figures 3 and 4, distributions of both datasets are compared. CPU and memory usage distributions of both datasets are similar. Google [12] and Alibaba [18] do not follow the nominal distribution. A right-skewed distribution can be observed as most of the tasks have very low CPU usage. Moving from average towards high usage, there are lesser number of tasks. This distribution can be more understandable using measure of center and measure of spread given in Table 2.

In cloud datasets, distribution is changing over a period of time which can be detected and visualized using a distance measure technique. KS statistical test is to detect changes in distribution. Kolmogorov–Smirnov (KS) test [1] is a non-parametric test suitable for cloud non-Gaussian distribution to confirm the existence of drift in current datasets. It is applied to the dataset and positive results confirmed the existence of drift. Drift magnitude is a property that can give a clear picture of the change. Drift magnitude can be visualized using

Table 2. Measure Center and Spread of the Distribution

DataSet		Alibaba		Google	
Resource Type		CPU	Memory	CPU	Memory
Measures of Center	Mean	0.029	0.004	0.016	0.021
	Median	0.011	0.001	0.002	0.006
	Mode	0.000	0.000	0.000	0.000
Measure of Spread	Standard Deviation	0.065	0.025	0.031	0.035
	Range	1	1	0.4736	0.752

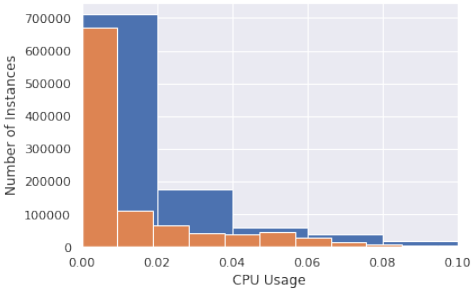

Fig. 3. CPU Usage Distribution in Google and Alibaba Usage

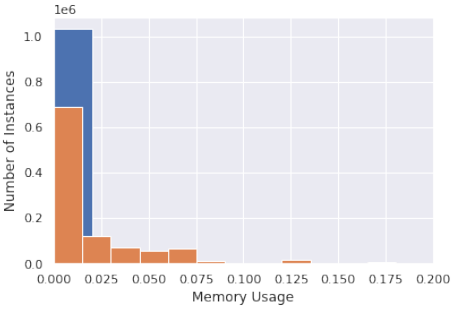

Fig. 4. Memory Usage Distribution in Google and Alibaba Usage

Hellinger Distance (HD), or Total Variation (TV) [17]. Drift detection follows the following major steps. We can process one or compare two windows to check the drift existence. In both cases, there will be a buffer to store some sample data to detect drift in the stream. Buffer data will be divided into required windows. These windows will contain data from equal amounts of data in sequence but from different time periods. HD is the most commonly used to measure distance. In this research, drift magnitude is calculated using the HD method as shown in Figure 4. Google usage trace shows a recursive pattern but that cannot be observed in the drift magnitude. Higher drift frequency is present in Google as compared to Alibaba as more spikes can be visualized in drift magnitude. Whereas, the Drift Magnitude of Alibaba is high.

In Figure 6, CPU and memory usage are compared against each other. An increase in CPU Usage has a direct effect on memory usage still there was some unusual presence of instances. As more instances of Google having high CPU and low memory usage. Alibaba also has instances of high memory usage and low CPU usage. To have better insight, the CPU and memory usage correlation is compared and visualized using correlation. We can see it does not follow a normal distribution. To calculate the correlation between CPU and Memory usage, Spearman and Kendalltau measures are used. This correlation test is non-parametric and designed for the non-normal distribution. Two consecutive windows from respected data are used to calculate the correlation. Using Spear-

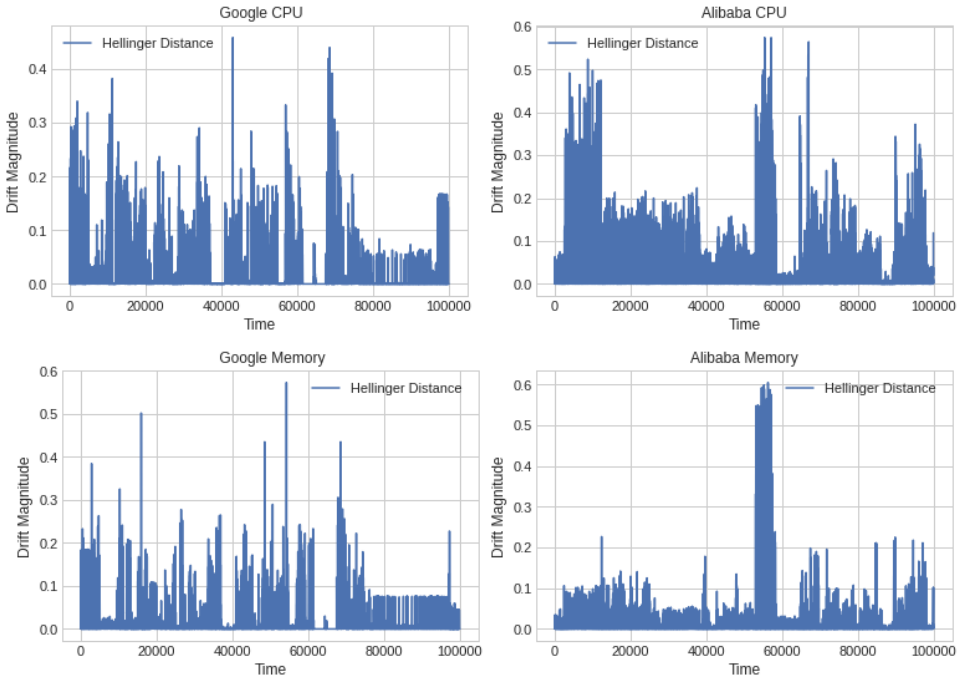

Fig. 5. Alibaba and Google Drift Magnitude

man's in Google and Alibaba resources shows a correlation of 0.77 and 0.64. Kendalltau shows a correlation of 0.54 for Google and 0.45 for Alibaba. These values of both the tests show a strong correlation between the two variables.

Drift detection and visualization need to divide data into sets of windows in order to compare. In this experiment, we have visualized two consecutive windows in clusters. Continuous numerical values of CPU and memory usage are difficult to understand. To compare two different datasets, we need to generalize the usage into clusters. As mentioned earlier, we can see the usage capacity level is defined as high, medium, and low. In Figure 7, we can see in windows 1 and 2 have nearly the same relation of variables but the cluster formation is different. These changes in formation show the drift at the cluster level. A bigger variation in cluster size can be seen in Alibaba, Figure 8. Alibaba set is showing a high drift magnitude as compared to google at Cluster Level. In case of more scattered relationships, a cluster level drift visualization is more useful. If data is less scattered than cluster-level this might not capture all the changes in clusters. Thus, simple instance-level drift visualization can do the job. We can see in the Figure 7 that Alibaba instances are very scattered to google. In the case of Alibaba spread is high thus these changes in clusters can be visualized more

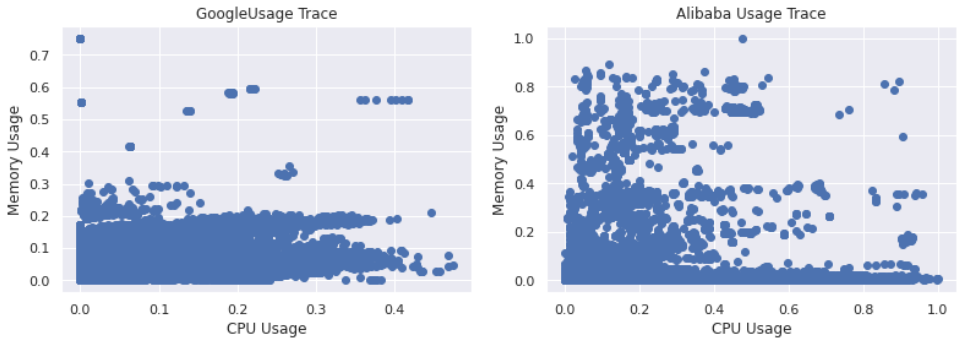

Fig. 6. Google And Alibaba Resources Correlation

clearly as compared to google even though both showed the drift at the cluster level.

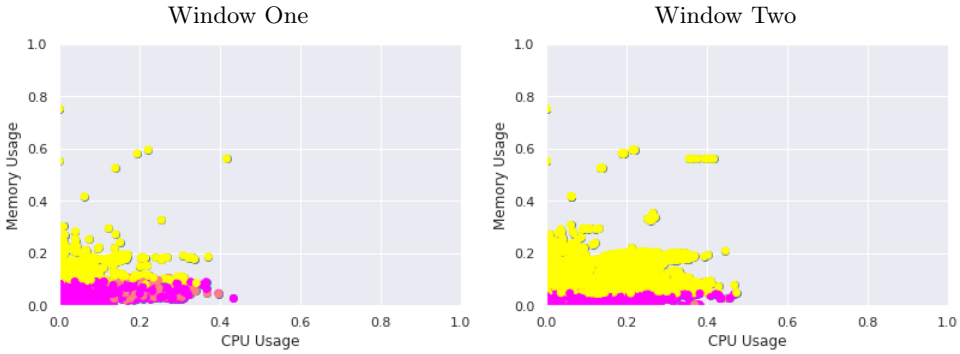

Fig. 7. Cluster Visualization of Google Resource Usage

4.1 External Factors

Multiple and heterogeneous types of users are using the cloud. The facilities cloud provider is claiming is also creating challenges. Cloud allows its users to scale up or down according to their requirements. This rapid elasticity causes more dynamic resource usage. The pay-as-you-go model is also allowing users to reduce costs but creates more variance in the usage. These variations in distribution are not only because of the Cloud providers. There are many external factors that contribute to these changes. Some factors can be personal to users.

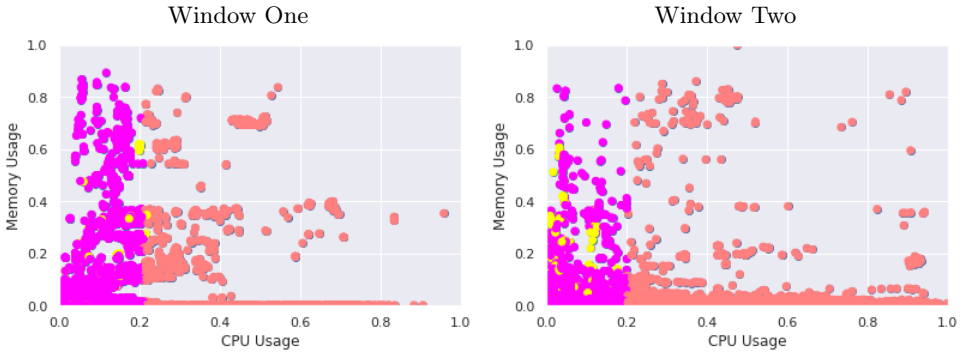

Fig. 8. Cluster Visualization of Alibaba Resource Usage

Epidemic, Seasonal, and Trend peaks can cause a sudden drift. Thus, Cloud has a fluctuating cloud users resource demand [5].

5 Conclusion and Future Work

Understanding CPU usage is essential for cloud providers. Visualization through most statistical and clustering methods is used to gain more insight and is important for resource prediction. Cloud characteristics, machine, and users heterogeneity, all factors are contributing to the drift. Continuous values are difficult to visualize. Thus, cluster-based visualization helps to see drift at the capacity group level. The above experimentation concludes that more scattered cloud usage data can be visualized more suitably at the cluster level. Alibaba and Google datasets are used to visualize changes in the capacity groups. In the future, this research can be applied in for real-time visualization and drift detection in the dynamic learning environment.

References

1. Berger, V.W., Zhou, Y.: Kolmogorov–smirnov test: Overview. Wiley statsref: Statistics reference online (2014)
2. Bholowalia, P., Kumar, A.: Ebk-means: A clustering technique based on elbow method and k-means in wsn. International Journal of Computer Applications **105**(9) (2014)
3. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) **41**(3), 1–58 (2009)
4. Inselberg, A.: Visualization and data mining of high-dimensional data. Chemometrics and intelligent laboratory systems **60**(1-2), 147–159 (2002)
5. Kaur, G., Bala, A., Chana, I.: An intelligent regressive ensemble approach for predicting resource usage in cloud computing. Journal of Parallel and Distributed Computing **123**, 1–12 (2019)

6. Kuyoro, S., Ibikunle, F., Awodele, O.: Cloud computing security issues and challenges. International Journal of Computer Networks (IJCN) **3**(5), 247–255 (2011)
7. Maaradji, A., Dumas, M., Rosa, M.L., Ostovar, A.: Detecting sudden and gradual drifts in business processes from execution traces. IEEE Transactions on Knowledge and Data Engineering **29**(10), 2140–2154 (2017). https://doi.org/10.1109/TKDE.2017.2720601
8. Nandgaonkar, S.V., Raut, A.: A comprehensive study on cloud computing. International Journal of Computer Science and Mobile Computing, a Monthly Journal of Computer Science and Information Technology **3**, 733–738 (2014)
9. Polyvyanyy, A.Y.C.D.C.J.M.A.: Visual drift detection for sequence data analysis of business processes. IEEE Transactions on Visualization and Computer Graphics (Early Access) pp. 1–1 (2018). https://doi.org/10.1109/TVCG.2021.3050071
10. Pratt, K.B., Tschapek, G.: Visualizing concept drift. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 735–740. KDD '03, Association for Computing Machinery, New York, NY, USA (2003). https://doi.org/10.1145/956750.956849, https://doi.org/10.1145/956750.956849
11. Reiss, C., Tumanov, A., Ganger, G.R., Katz, R.H., Kozuch, M.A.: Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In: Proceedings of the third ACM symposium on cloud computing. pp. 1–13 (2012)
12. Reiss, C., Tumanov, A., Ganger, G.R., Katz, R.H., Kozuch, M.A.: Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In: Proceedings of the Third ACM Symposium on Cloud Computing. SoCC '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2391229.2391236, https://doi.org/10.1145/2391229.2391236
13. Stiglic, G., Kokol, P.: Interpretability of sudden concept drift in medical informatics domain. In: 2011 IEEE 11th International Conference on Data Mining Workshops. pp. 609–613 (2011). https://doi.org/10.1109/ICDMW.2011.104
14. Syakur, M., Khotimah, B., Rochman, E., Satoto, B.D.: Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In: IOP conference series: materials science and engineering. vol. 336, p. 012017. IOP Publishing (2018)
15. Wang, X., Chen, W., Xia, J., Chen, Z., Xu, D., Wu, X., Xu, M., Schreck, T.: Conceptexplorer: Visual analysis of concept drifts in multi-source time-series data. In: 2020 IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 1–11. IEEE (2020)
16. Webb, G.I., Lee, L.K., Goethals, B., Petitjean, F.: Analyzing concept drift and shift from sample data. Data Mining and Knowledge Discovery **32**(5), 1179–1199 (2018)
17. Webb, G.I., Lee, L.K., Petitjean, F., Goethals, B.: Understanding concept drift. arXiv preprint arXiv:1704.00362 (2017)
18. Weng, Q., Xiao, W., Yu, Y., Wang, W., Wang, C., He, J., Li, Y., Zhang, L., Lin, W., Ding, Y.: {MLaaS} in the wild: Workload analysis and scheduling in {Large-Scale} heterogeneous {GPU} clusters. In: 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22). pp. 945–960 (2022)
19. Yang, C., Huang, Q., Li, Z., Liu, K., Hu, F.: Big data and cloud computing: innovation opportunities and challenges. International Journal of Digital Earth **10**(1), 13–53 (2017)