Evaluation of Concept Drift in Poisson Big Data Stream using Adaptive Sliding Windows

Chanintorn Jittawiriyanukoon* and Vilasinee Srisarkun**

Abstract

Involving with big data whose dynamically changes over time is one of the major problems in big data curation. In this research an adaptive sliding window will be presented, as an evaluation with memory and variable length of data stream. Adaptive sliding window concept is to maintain the variable window size in order to carry the latest read data stream arriving in Poisson process from which older ones based upon algorithm rules. We need to involve the change of concepts meaning (i.e. concept drift) which is necessary for data releases and sophisticate data links. The concept drift thus reflects the change of window size and provides statistics update from recent data. Our simulation runs both fixed and continuous data stream so that sliding window is applied to different processing of data curation. In this paper we have proposed Poisson and Random arrival model of data stream which will employ Massive Online Analysis (MOA) for evaluating the concept drift measurements. Stagger stream generator with the Hoeffding bound outperforms and results highest accuracy while Naïve Bayes learner with Gradually Change generator fits Poisson arrival pattern.

Keywords: Big Data Curation, Concept Drift, Poisson Process, Sliding Windows, MOA

* Graduate School of eLearning, Assumption University
88 Moo 8, Bang Na Trad Km 26, Bang Sao Thong District, Samut Prakan 10540, THAILAND.
E-mail: chanintornjtt@au.edu

** Martin de Tours School of Management and Economics, Assumption University
88 Moo 8, Bang Na Trad Km 26, Bang Sao Thong District, Samut Prakan 10540, THAILAND.
E-mail: vilasineesrs@au.edu

การประเมินผลการเปลี่ยนแปลงรูปแบบของข้อมูลขนาดใหญ่ที่เข้ามาตามการกระจายปัวซองโดยใช้การปรับเปลี่ยนขนาดหน้าต่าง

ชนินทร จิตตวิริยานุกูล* และ วิลาสิณี ศรีศกุน **

บทคัดย่อ

ข้อมูลขนาดใหญ่ที่มีรูปแบบเปลี่ยนแปลงแบบไดนามิกในช่วงเวลาใดเวลาหนึ่งมักก่อให้เกิดปัญหาสำคัญในการดูแลจัดการข้อมูลที่มีขนาดใหญ่นั้นเสมอ งานวิจัยนี้ใช้การปรับค่าหน้าต่างเพื่อรองรับและประเมินผลของข้อมูลที่มีรูปแบบเปลี่ยนแปลงแบบไดนามิกว่าจะมีผลต่อหน่วยความจำมากน้อยเพียงใด การปรับเปลี่ยนค่าหน้าต่างเพื่อรักษาสมดุลขนาดของหน้าต่างให้เหมาะสมกับการดำเนินการกับสตรีมข้อมูลล่าสุดที่มีรูปแบบการเข้ามาสัมพันธ์กับการกระจายแบบปัวซอง ทั้งนี้ขึ้นอยู่กับกฎเกณฑ์และขอบเขตที่กำหนดไว้ในแต่ละอัลกอริธึมนั้น ๆ เราจำเป็นที่ต้องศึกษาการเปลี่ยนรูปแบบของข้อมูล (Concept Drift) ซึ่งเป็นสิ่งจำเป็นสำหรับการจัดเก็บ เผยแพร่ข้อมูลรวมถึงการเชื่อมโยงข้อมูลไปยังส่วนอื่น ๆ การเปลี่ยนรูปแบบของข้อมูลนี้จะมีผลโดยตรงกับการปรับเปลี่ยนขนาดของหน้าต่างและการปรับค่าสถิติต่างๆโดยคำนวณได้จากข้อมูลล่าสุด การจำลองการทำงานดังกล่าวทั้งสำหรับข้อมูลรูปแบบคงที่และข้อมูลที่มีรูปแบบปลี่ยนแปลงแบบไดนามิกและต่อเนื่อง การปรับค่าหน้าต่างถูกนำมาประยุกต์ใช้ในการประมวลผลและจัดการข้อมูลดังกล่าว งานวิจัยนี้ได้นำเสนอข้อมูลที่มีรูปแบบการเข้ามาแบบการกระจายปัวซองและการกระจายแบบสุ่ม นำมาวิเคราะห์หาประสิทธิภาพโดยใช้เครื่องมือที่เรียกว่า MOA (Massive Online Analysis) ผลการวิเคราะห์ พบว่า Stagger อัลกอริธึมมีประสิทธิภาพเหนือกว่าและมีความแม่นยำสูงสุดเมื่อนำมาใช้ร่วมกับขอบเขตของ Hoeffding ส่วนขอบเขตของ Bayes นั้นเหมาะกับรูปแบบการเข้ามาของข้อมูลที่มีการกระจายแบบปัวซองมากกว่า

คำสำคัญ: การจัดการข้อมูลขนาดใหญ่ การเปลี่ยนรูปแบบของข้อมูล การกระจายปัวซอง การปรับค่าหน้าต่าง

* บัณฑิตวิทยาลัยการศึกษาทางอิเล็กทรอนิกส์ มหาวิทยาลัยอัสสัมชัญ
เลขที่ 88 หมู่ 8 ถนนบางนา-ตราด กม.26 เขตบางเสาธง จังหวัดสมุทรปราการ 10540
อีเมล: chanintornjtt@au.edu

** คณะการจัดการและ เศรษฐศาสตร์ มาร์ติน เดอ ตูร์ มหาวิทยาลัยอัสสัมชัญ
เลขที่ 88 หมู่ 8 ถนนบางนา-ตราด กม.26 เขตบางเสาธง จังหวัดสมุทรปราการ 10540
อีเมล: vilasineesrs@au.edu

Introduction

Currently huge volume of data streams are fluctuated from real time environment, computer networks, Internet traffic, social media traffic etc. Different types and unstructured data traverse continuously at fast speed with varying arrival rate. Big data generator has different patterns, including volume, velocity and variety. Selecting a processing environment and curating for big data is a critical task as it would be a proper solution for volume, velocity and variety of data. The "volume" refers to the amount of incoming data, "velocity" refers to high speed arrival rate and "variety" refers to data types (structured and unstructured data). The structured one can be databases, documents while the unstructured data will be in the form of email, tweets, images, videos and audio files. Due to this variety of unstructured data, the problems of storage management, curation and data analysis arise. Big data can be characterized by velocity which is the rate of data arrival (rate of data flows from sources) especially from social media sites, mobile communication networks, etc.

Adaptive sliding window (ASW) plays a critical role in the analysis of concept drift. Concepts are central entities in the system and represent objects with common characteristics. However, with time, objects are constantly subject to change. In other words, concepts are naturally influenced by concept drift, the change of their meaning over time (Bose et al., 2013). Big Data contains loosely structured information of great business value thus if we can extract those valuable insights we can make much better decisions for the market research and strategy. An outlier detection in some instances will extract a set of inconsistent data from the whole sources (V, J. Hodge, 2014). Alternative method which is simple and straightforward based upon "map" and "reduce" functions from LISP programming language is called Hadoop (https://www.ibm.com). Apache Hadoop for instance is an open-source module supporting distributed shuffles. It is beneficial for the distributed shuffle operation (which cuts communication cost) and fault tolerance characteristics of the MapReduce model.

The aim of this research is to evaluate the concept drift of big data with Poisson arrival process using ASW. There are some challenges for manipulating big data. Firstly, the data arrives at high speed then rescanning the whole database is nearly impossible in practice. Secondly, limited space is available to store big data for online processing. The curation needs to adapt to this challenge. Lastly, curation algorithm must be faster than

the arrival rate to avoid a bottleneck while manipulating. Approximation method can be an alternative to help fasten the manipulation if the accuracy of results is trivial.

Concept Drift

A. Data Stream Applications

Present research in big data is focused to dynamic situations, where formats hidden in data structure are varied and each data tuple can be reached more than once. The most information structure is arrangement (classification), characterized as a known structure to apply configuration to another information. These techniques result in environments to process massive data streams and their different patterns - concept drift. Concept drift is a generic name used for changes over the time in the machine learning structure. Changes associate replacements of one order with another, but also are inclusive of steady trends and minor shifts of the essential probability distributions specified in each algorithm (Ludmila, I. Kuncheva, 2004). For most classifiers the incident of concept drift induces to a severe drop in classification accuracy. This is the reason why, new data manipulation involved with classifiers to help improve classification accuracy dedicated to data streams has been proposed.

B. Adaptive Sliding Windows (ASW)

The motivation of concept drift in data streams has introduced sliding-window approaches which lead to a dismissing process, which limits the number of processed data in order to develop changes in classification. Different approaches to manipulating data streams include drift detection, bagging, twitter stream, ensemble classifiers, tree generator and using Hoeffding bound to evaluate classification accuracy and performance. A simulation framework called Massive Online Analysis (MOA) for these algorithms and applying for big data streams has been developed by (Albert, Bifet, Geoff, Holmes, Richard, Kirkby & Bernhard, Pfahringer, 2010). It includes both offline and online data curation approaches as well as analysis tools for their performance evaluation. MOA is a tool that can simulate and run experiments on the development of new time-varying data stream classification. Time consumption, memory costs and classification's accuracy of ASW algorithm with synthetic data will be collected.

ASW strategies involve divergence of the sliding window approach. A window size is preserved that keeps the latest read examples, and from which older examples are decreased regarding to set of rules provided by algorithm. The contents of the window can be purposely used for a) to detect change b) to develop statistics from the latest examples, and c) to generate a future revised model afterwards.

ASW has been introduced to use windows of variable size (sliding windows). In general, in case of known bound on the rate of change, ASW can perform an adaptive size for window. Hoeffding's bound is used to guarantee a streaming algorithm. ASW is popularly used in various time-varying applications for instance in the medical analysis, weather forecast, social media context analysis and financial area. It varies the size of the window and example segment increases until the error of the segment reaches the approximate current segment and specific threshold. More specifically, an established window is dropped if there is sufficient sign that its average size contrasts from that of the remaining window. This leads two significances: first, that change dependably declared whatever point the size of window decreases; and second, that the mean over the current window can be dependably taken as an average in the present stream. ASW does not keep the window size clearly, but shrinks it using a variance of the exponential value in the histogram. That means it maintains a window size (*W)* using only O (log*W*) memory space and O (log*W*) manipulating time per instance.

C. Bayesian Classifier

The classifier designed by MOA employs traditional Bayesian evaluation while assuming that all inputs are naive and autonomous (Albert, Bifet, Geoff, Holmes, Richard, Kirkby & Bernhard, Pfahringer, 2010). Naïve Bayes is a straightforward classifier algorithm with low cost in computation. In different classes, the prepared Naïve Bayes classifier will forecast every unlabelled instances of each class with high accuracy.

D. Hoeffding Tree Algorithm

It is one of the decision tree induction algorithms for data stream. A Hoeffding tree (Albert, Bifet, Eibe, Frank, Geoffrey, Holmes & Bernard, Pfahringer, 2007). is an incremental algorithm which is taking from enormous data streams, assuming that the distribution producing examples will not diverge over time. Hoeffding tree performs that an insignificant sample can be adequate to select an intense attribute.

E. Concept Generator Functions

Streaming Ensemble Algorithm (SEA) generates concept functions. This dataset contains abrupt concept drift as listed in (W.N. Street. & Y, Kim., 2001), (Wang, H., Fan, W., Yu, P.S. & Han, J., 2003). Concept functions are based on three attributes, where the first two attributes (f_1 and f_2) are applicable. Three attributes have figures ranging from 0 to 10. The dataset are distributed into 4 blocks of dissimilar concepts. The classification of each block is developed using $f_1 + f_2 \leq \alpha$ where f_1 and f_2 denote the 1st and the 2nd attributes and α is a threshold point. The STAGGER generates concept functions as introduced in (W.N. Street. & Y, Kim., 2001), (J.C. Schlimmer. & R.H. Granger, 1986). Concepts are based on Boolean functions of 3 attributes: size (small, medium, or large), shape (triangle, circle, or rectangle), and color (green, red, or blue). For example a concept attribute binding either blue circles or green rectangles will be denoted by (shape=circle and color=blue) or (shape=rectangle and color=green). Hyperplane generates concept with a problem of predicting class of a rotating plane (G, Hulten, L, Spencer & P, Domingos, 2001), (Cunningham, P., Nowlan, N., Delany, S.J. & Haahr, M., 2003). It is possible for running time-varying concepts as we can alter the alignment and location of the plane easily by altering the comparative size of the weights. Concept drift is led to this dataset taking drift to each weight attribute $w_i = w_i + \Delta\sigma$, where σ is the possibility that the direction of alteration is opposite and Δ is the alteration applied to every instance. LED generates concept with a problem of predicting the 7-segment-LED-digit display, where each attribute has a small possibility of being reversed (Cunningham, P., Nowlan, N., Delany, S.J. & Haahr, M., 2003). The generator used for LED experiments develops 24 binary attributes while 17 of which are extraneous. It is used to collect a million instances with gradual or abrupt concept drift. Random Radial Basis Function (RandomRBF) (Cunningham, P., Nowlan, N., Delany, S.J. & Haahr, M., 2003) will generate a stream offering a complicated type of concept that is devious to estimate with a decision tree option. The RBF produces functions as follows: A random centroid is produced. Each center has a class label, a standard deviation, random position and class weight. New examples are created by randomly choosing a center, taking weights into account so that centers with bigger weight are to be adopted. A random path is selected to compensate the attribute from the center. The displacement length is randomly placed from a normal distribution with standard deviation regulated by the selected centroid. Only numeric attributes are generated from the algorithm. Lastly, Waveform (Cunningham, P., Nowlan, N., Delany, S.J. & Haahr, M., 2003) generates a concept with a problem of foretelling one of 3 waveforms. Each of which is produced from a combination of 2 or 3 fundamental waves. There are 2

types, wave21 which reflects 21 arithmetic attributes and wave40 which leads another 19 extraneous attributes.

F. Stream Drift Analysis

Concepts refer to the target variables, in which the model is trying to forecast. Concept change is the alteration of the basic concept over simulation-run-time. Concept drift outlines a gradual alteration of the concept as concept changes slowly and gradually. However, whenever concept shift arises and causes difference between two concepts, this refers to abrupt change. It is also known as a sampling alteration or shift concept drift. Even though the concept remains unchanged, the alteration will induce to adjusting the current model as the model's error amount may not be tolerable with the new concept.

Poisson Data Stream

Poisson traffic actually will work with as real-time multimedia applications in practice. If receiving any emails doesn't affect the arrival times of future arrival of emails, i.e., if emails from any senders arrive independently of one another, then a practical assumption is that the number of emails received in an hour follows a Poisson distribution. In implementing and designing multimedia over clouds, Poisson model has been widely implemented. In data stream environment, a new kind of Poisson arrival process could be adopted for real-time processing applications. The number of incoming stream received by a cloud per minute obeys Poisson distribution for instance. Instead of Poisson applications, traffic with Erlang distribution for multimedia can also be found in (C Jittawiriyanukoon, 2014), (Koo et al., 1999). However, traffics in this paper include voice, video and data stream are analyzed and considered to ensure the concept drift as described in previous section. The stream is sorted out by the sequence of data entering hence Poisson formula is applied for data arrival process. Multimedia traffic over clouds had been discussed in (Amreen, Khan & Kamal, K. Ahirwar, 2011).The Poisson distribution is a continuous probability density function with wide utilization due to its association to the exponential function. The distribution is also applicable in the fields of random processes and radioactive sources. The probability density function (pdf) of the Poisson distribution is

$$f(k,\lambda)=\lambda^{k}e^{-\lambda}/k! \quad \text{for} \quad \lambda \geq 0 \tag{1}$$

where k is a positive integer and is also called the shape parameter, and λ is the arrive rate as shown in Equation 1.

Results and Analysis

The experiment in this paper is categorized into four parts which include classifiers, learner evaluation, data stream generation, and performance measurement. MOA also allows user to create dynamic data streams using algorithm in generators, by connecting new generated streams, or discarding streams. MOA also provides a characteristic that allows to enumerate concept drift to any data streams. The sample of MOA configuration setting is shown in Table 1.

Table 1: MOA Configuration Setting

Classification	**Evaluate Prequential**
Learner	NaiveBayes (NB) and HoeffdingTree (HT)
Stream	SEA, STAGGER, HYPER PLANE, LED,
Generator	RANDOMRBF, WAVEFORM
Evaluator	PerformanceEvaluator
Performance	ADWIN
DriftStream	ConceptDriftStream (CD)

The experiment of data stream generators regarding to NB and Hoeffding classifiers with concept drift (CD, LED, RANDOMRBF, WAVEFORM, SEA, STAGGER, HYPERPLANE) and data stream records for Poisson arrival and Random arrival time. The MOA for concept drift configuration is listed in Table 2.

Table 2: MOA Setting for Concept Drift

Classification	**Concept DriftStream**
Learner	AdwinChangeDetector
Stream	Poisson Arrival (PA) and Random
	Arrival (RND) Streams
Evaluator	AdwinPerformanceEvaluator
Performance	ADWIN0 with Poisson Process
StreamProblem	GradualChangeGen and
	AbruptChangeGen

The experimental simulation set up is as of subsequent parameters. Instance-limit: 10^8 instances, PerformanceEvaluator: ADWIN0, Classification: Evaluate Prequential, Classifier: NaiveBayes and Hoeffding. The synthetic data set (arff file) from a company consists of

name, surname, ID, email account, IP address and arrival time attributes. This is a dataset with four-thousand records (320 KB). The single input data stream is considered in this paper while results of MOA simulation for NaiveBayes (NB) for static data streams with concept drift can be found in (Srimani & Patil, 2016).

Table 3, 4, 5 and 6 will demonstrate the performance metrics, the accuracy, kappa, ram-hrs, utilized time and memory space consumption for each experiment of data stream generators regarding to NB and Hoeffding classifiers with concept drift (CD, RANDOMRBF, LED, and WAVEFORM), by combining concept drift (STAGGER, SEA, and HYPERPLANE) and data stream records for Poisson arrival and Random arrival time.

It is found that the NB algorithm's performance outperforms on STAGGER generator with accuracy = 100%, Kappa = 100 when compared to others as shown in Table 3. The concept drift is supplemented to all experimental generators i.e. CD, STAGGER, SEA, and HYPERPLANE by using stream of concept drift available from MOA simulation. The constant results of ram-hours and memory space consumption are distinctively monitored in this experiment. Note that the classifier somehow is telling us something practical about the concept of experimental data, and that is as well involving part of machine learning. As results of best accuracy 100% attained above, it is not surprising figure rather they are importing the robust classifier performance of a concept drift to an automated system in the future.

Table 3: Results of Evaluation Measures for NB

NB	CD	SEA	STAGGER	HYPERPLANE
Accuracy (%)	73.6	88.2	100	94.1
Kappa	44.4	73.3	100	88.3
Kappa Temp	45.9	74.3	100	88.3
Ram-Hrs	0	0	0	0
Time (sec)	20	21	20	20.8
Memory (MB)	0	0	0	0

Table 4 shows that it is apparent that the NB algorithm's performance is pleasurable on WAVEFORM generator with accuracy = 80.4%, Kappa = 70.6 after comparing to others. The RANDOMRBF, LED and WAVEFORM generators used in the experiment are time-varying

with concept drift streams in MOA simulation. The constant results of ram-hours and memory space consumption are distinctively monitored in this experiment. It is noted that it is uneasy figure of memory space especially for LED and WAVEFORM.

Table 4: Results from Simulation for NB with Convert Drift

NB	LED	RANDOMRBF	WAVEFORM
Accuracy (%)	73.9	71.9	80.4
Kappa	71	43.9	70.6
Kappa Temp	71	43.9	70.6
Ram-Hrs	0	0	0
Time (sec)	20.2	20.7	20.7
Memory (MB)	0.02	0	0.01

It is interesting that the HT algorithm's performance also outperforms on STAGGER generator with accuracy = 100%, Kappa = 100 as compared to others as shown in Table 5. The concept drift is enhanced to all experimental generators i.e. CD, SEA, HYPERPLANE and STAGGER by employing concept drift stream available from MOA simulation. The constant results of ram-hours are distinctively monitored in the present experiment. CD consumes highest memory space = 4.48MB as compared to others. Similarly, notice that the classifier somehow is showing something practical about the concept of experimental data which is reflecting a machine learning. Accuracy 100% attained above is no longer unexpected results but they are transporting the vigorous classifier performance of a concept drift to automation rather. In addition, the error rate of misclassification is assumed by (Albert, Bifet., Geoff, Holmes., Richard, Kirkby. & Bernhard, Pfahringer., 2010) in which the error rate of the machine learning algorithm will decline once the number of samples hikes provided that the distribution function of the samples is immobility. This research focuses on big data directing to a huge number of samples then the error rate of stream generator can be notably neglected.

Table 5: Results of Evaluation Measures for HT

HT	CD	SEA	STAGGER	HYPERPLANE
Accuracy (%)	97.6	89.7	100	90.6
Kappa	95.1	77.1	100	81.3
Kappa Temp	95.1	77.6	100	81.3
Ram-Hrs	0	0	0	0
Time (sec)	20.3	20.1	20.3	20.2
Memory (MB)	4.48	1.92	0	3.43

Table 6 demonstrates that it is obvious that the HT algorithm's performance is superior on RANDOMRBF generator with accuracy = 93%, Kappa = 86 as compared to others. The RANDOMRBF, LED and WAVEFOR generators are employed in the experiment are time-varying with concept drift streams in MOA simulation. The constant results of ram-hours are distinctively monitored in this experiment. It is noted that it is uneasy figure of memory space especially for all these LED, RANDOMRBF, and WAVEFORM.

Table 6: Results of Evaluation Measures for HT With Concert Drift

HT	LED	RANDOMRBF	WAVEFORM
Accuracy (%)	73.9	93	84.3
Kappa	71	86	76.5
Kappa Temp	70.9	86	76.6
Ram-Hrs	0	0	0
Time (sec)	20.4	20	20.4
Memory (MB)	0.67	3.80	2.10

It is found that the NB algorithm's performance is excellent on Poisson arrival process with Gradual Change Generator (time = 60 msec) when compared to Random arrival process (time = 90 msec) as shown in Table 7. It is noted that the NB algorithm's performance is agreeable on Random arrival process with Abrupt Change Generator (time = 60 msec) when compared to Poisson arrival process (time = 80 msec). The constant results of memory space consumption (10 KB) are distinctively detected in the present experiment. MOA simulation allows effectively assess classifier algorithms on hefty data streams, in the sequence of several millions of samples, and under conservative memory usage. The figure shows maximum memory bound used by the algorithm in order to execute individual dataset.

Table 7: Results of Evaluation Measures for Poisson Arrival Process With NB Learining Slgorithm

NB	Grad		Abrupt	
	PS	RND	PS	RND
Time (msec)	60	90	80	60
Memory (KB)	10	10	10	10

Table 8 lists that it is apparent that the HT algorithm's performance is qualified on Random arrival process with Gradual Change Generator (time = 30 msec) when compared to Poisson arrival process (time = 60 msec). It is to be noted that the HT algorithm's performance is also fine on Random arrival process with Abrupt Change Generator (time = 60 msec) when compared to Poisson arrival process (time = 80 msec). The constant results of zero memory space consumption are distinctively detected in the present experiment since ASW will not retain the window size explicitly, but shortens it using a technique of a variance of the exponential history. It thus explains it maintains a window size (W) consuming only little memory space (next to zero) as specified in section B previously.

Table 8: Results of Evaluation Measures for Poisson Arrival Process with HT Learining Algorithm

HT	Grad		Abrupt	
	PS	RND	PS	RND
Time (msec)	60	30	80	60
Memory (KB)	0	0	0	0

Figure 1 represents the graph generated for the accuracy valuation results collected from NB and HT algorithms. It is clearly demonstrated that the performance of both HT and NB algorithms is superior on STAGGER generator with accuracy = 100%, Kappa= 100 as compared to others. HT algorithm outperforms on any generators as compared to NB algorithm. ASW will not retain the window size explicitly, but shortens it using a technique of a variance of the exponential history. It thus explains it maintains a window size (W) consuming only little memory space (next to zero) as specified in section B previously.

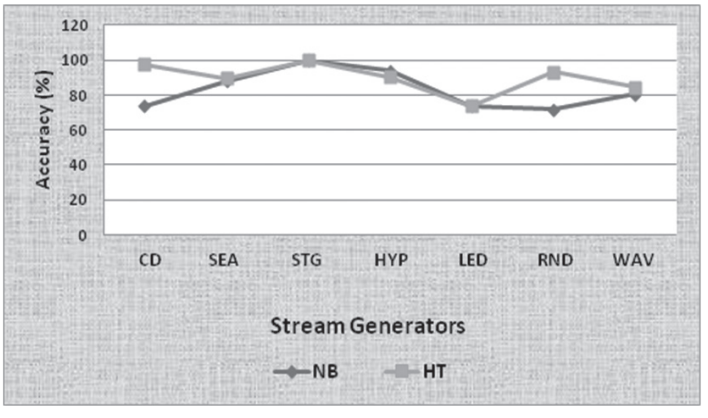

Fig.ure 1: Graph of Accuracy Evaluation for Data Streams with NB and HT Algorithms

Conclusion

The commitments of the research paper are as follows. The research involves concept drift in both static and time-varying data streams in MOA simulation. The classification employing Bayesian and Hoeffding bound is applied for investigating concept drifts performance evaluation. The synthetic data set with Poisson arrival process is taken into account. Different classifiers and generator algorithms are observed as concept drifts, which are included during the preparation of the simulation model. Regarding to the MOA results, the problems of data streams such as the amount of memory space consumption, ram-hours and execution time are collected to the performance evaluation centric. The paper concludes as follows. In the investigation authors have employed data set with Poisson and Random arrival time. These data sets will be simulated with 7 generators which are 4 are static streams and 3 time-varying streams to perform big data curation with concept drift employing MOA. The learning algorithms adopt Baysian and Hoeffding Tree approaches. CD, STAGGER, SEA, and HYPERPLANE are 4 static data streams for which real-stream-concept-drift is enhanced. WAVEFORM, RANDOMRBF, and LED are 3 time-varying data set generators which are developed with concept drift during simulating. The NB algorithm's performance outperforms on STAGGER generator with accuracy = 100%, Kappa = 100 as compared to others. HT algorithm is also proven to be superior on STAGGER generator with accuracy = 100%, Kappa = 100 as compared to others. The difference is that HT algorithm will consume higher memory space as compared to NB algorithm usage. It is apparent that the only NB algorithm's performance is excellent on Poisson arrival process with Gradual Change Generator when compared to Random arrival process. The distinctive constant figures of ram-hours and memory space are collected in this work. The future research will include the investigation of Erlang data stream using both Naïve Bayes and Hoeffding Tree approaches in MOA. Various companies such Amazon, Yahoo, Facebook etc. are adopting Erlang distribution in their manufacturing systems. For instance, Amazon opts Erlang to implement database services as a component of the Amazon Elastic Compute Cloud (EC2).

References

Albert, Bifet, Eibe, Frank, Geoffrey, Holmes & Bernard, Pfahringer. (2007). Accurate Ensembles for Data Streams Combining Restricted Hoeffding Trees Using Stacking. *Journal of Machine Learning Research*, 225-240.

Albert, Bifet, Geoff, Holmes, Richard, Kirkby & Bernhard, Pfahringer. (2010). MOA: Massive Online Analysis. *Journal of Machine Learning Research 11*, 1601-1604.

Amreen, Khan & Kamal, K. Ahirwar. (2011). Mobile Cloud Computing as a Future of Mobile Multimedia Database. *International Journal of Computer Science and Communication*, *2*(1): 219-221.

Bose et al. (2013). Dealing With Concept Drifts in Process Mining. *IEEE Transactions on Neural Networks and Learning Systems*, 1-18. (DOI: 10.1109/TNNLS.2013.2278313)

C, Jittawiriyanukoon. (2014). Performance evaluation of reliable data scheduling for Erlang multimedia in cloud computing. *Ninth International Conference on Digital Information Management (ICDIM)*, 39-44. (DOI: 10.1109/ICDIM.2014.6991394)

Cunningham, P., Nowlan, N., Delany, S.J. & Haahr, M. (2003). A Case-Based Approach to Spam Filtering that Can Track Concept Drift. *Proceedings of ICCBR*, Workshop on Long-Lived CBR Systems.

G, Hulten., L, Spencer. & P, Domingos. (2001). Mining Time-Changing Data Streams. *ACM Press*, San Francisco, CA, 97-106. https://www.ibm.com

J.C. Schlimmer. & R.H. Granger. (1986). Incremental Learning from Noisy Data. *Machine Learning*, *1*(3): 317-354.

Koo et al. (1999). Analysis of Erlang Capacity for the Multimedia DS-CDMA Systems. *IEICE Transaction Fundamentals*, E82-A(5): 849-855.

Ludmila, I. Kuncheva. (2004). Classifier Ensembles for Changing Environments. *Lecture Notes in Computer Science*, Springer, 1–15.

Srimani & Patil. (2016). Mining Data Streams with Concept Drift in Massive Online Analysis Frame Work. *WSEAS Transaction on Computers*, 15: 133-142.

Victoria, J. Hodge. (2014). Outlier Detection in Big Data. *IGI Global*, 1762-1771. (DOI: 10.4018/978-1-4666-5202-6.ch157)

W, N. Street. & Y, Kim. (2001). A Streaming Ensemble Algorithm for Large-Scale Classification. *Proceeding of 7th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, ACM Press, New York, USA, 377-382.

Wang, H., Fan, W., Yu, P.S. & Han, J. (2003). Mining Concept-Drifting Data Streams using Ensemble Classifiers. *9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining KDD*, ACM Press, 226-235.