Analyzing Performance of Classification Algorithms on Concept Drifted Data Streams

Aradhana Nyati
Sir Padampat Singhania University,
India

Divya Bhatnagar
Sir Padampat Singhania University,
India

Avinash Panwar
Sir Padampat Singhania University,
India

ABSTRACT
Current research in data mining concentrates on the development of new techniques for mining high-speed data streams. The fundamental data generation mechanism changes over the time, this is common in most real-world data streams, which introduces concept drift into the data. Mobile devices, streaming, remote sensing applications which are networked digital information systems, encounter the issue of the size of data and the capacity to be adaptive to changes in concept in real-time. In this paper the main issue of concept drift is addressed with real and synthetic data streams and the comparison of ensemble classifiers has been made in view of concept drift for the assessment of the performance. Various classifiers were applied on data stream with and without concept drift for analysis. This has resulted in better performance of the classifiers on the type of data whether it is categorical, numeric or alphanumeric.

Keywords
Data mining, Data Stream, Concept Drift, Classification

1. INTRODUCTION
Data stream is a continuous and changing sequence of data that continuously arrive at a system to store or process. It is vital to find out useful information from large enormous amount of data streams generated from different applications viz. organization record, call center record, sensor data, network traffic, web searches etc. One of the challenge of data streams are concept drift. Streaming data poses additional challenges for active learning, since the data distribution may change over time (concept drift) and classifiers need to adapt.

Concept drift causes problems because the learning become less accurate as time passes. If changes do not occur close to the boundary, they will be missed and classifiers will fail to adapt. In the situation of concept drift, a learning algorithm is required that can, detect concept changes, quickly recover from a concept change, adjust its hypotheses to a new context, make use of previous experience in situations where concept reoccurring happens. These data streams need to be analyzed for finding patterns which help us in segregating anomalies and forecasting future behavior. So classification is required. The classification model is a representation of classification rules, decision trees, neural networks, or mathematical computation which is used for classification. The objective of data stream classification is to predict the (categorical) class labels of a given data tuples based on a training data set and to develop a model of classifier.

2. RELATED WORK
Kadwe and Suryawanshi discussed various techniques to manage concept drifts. The synthetic and real datasets with different concept drifts and the applications are discussed [1]. Gama et al. defined adaptive learning process and categorization for handling concept drift and presented a set of illustrative applications [2]. Mittal and Kashyap suggested various online methods of drift detection in his paper. They presented results of experiments and comparison of online drift detection methods [3]. Bifet et al. proposed a new experimental framework for evaluating change detection methods against intended outcomes. They proposed framework could be used with other data mining tasks such as frequent item and pattern mining, clustering etc. [4]. CVDFT algorithm based on sliding time window is described by Hoeglinger et al. defined the comparison of experimental results to show the outperformance of CBDT which has been introduced [5]. Bifet et al. in their paper introduced evaluation methodology for big data streams which methodology addresses unbalanced data streams, data where change occurs on different time scales, and the question of how to split the data between training and testing, over multiple models [6]. Wankhade and Dongre introduced adaptive ensemble boosting approach for the classification of streaming data with concept drift. They used adaptive ensemble boosting method with the use of adaptive sliding window and Hoeffding tree [7].

Though several efforts have been made to overcome the concept drift challenge, the effects on the performance of classifier in the presence of concept drift in the data stream could not be clearly understood. In this work an effort has been made to analyze the behavior and performance of classification algorithm with concept drifted categorical, numeric and alpha-numeric data streams.

3. FRAMEWORK AND MEASURES

3.1 Framework
Fig. 1 shows the framework of the proposed process. Data streams are collected through data stream generator or collection of data stream. Various approaches are applied on data stream in MOA. Segment of data stream is selected for preprocessing. There will be two cases. First, if data stream contain concept drift then it will detect through MOA. After that classification is performed by different classifiers. Another one is, if data stream is without concept drift, then the classification algorithm is directly applied to data stream. After the classification process is complete, knowledge will be produced.

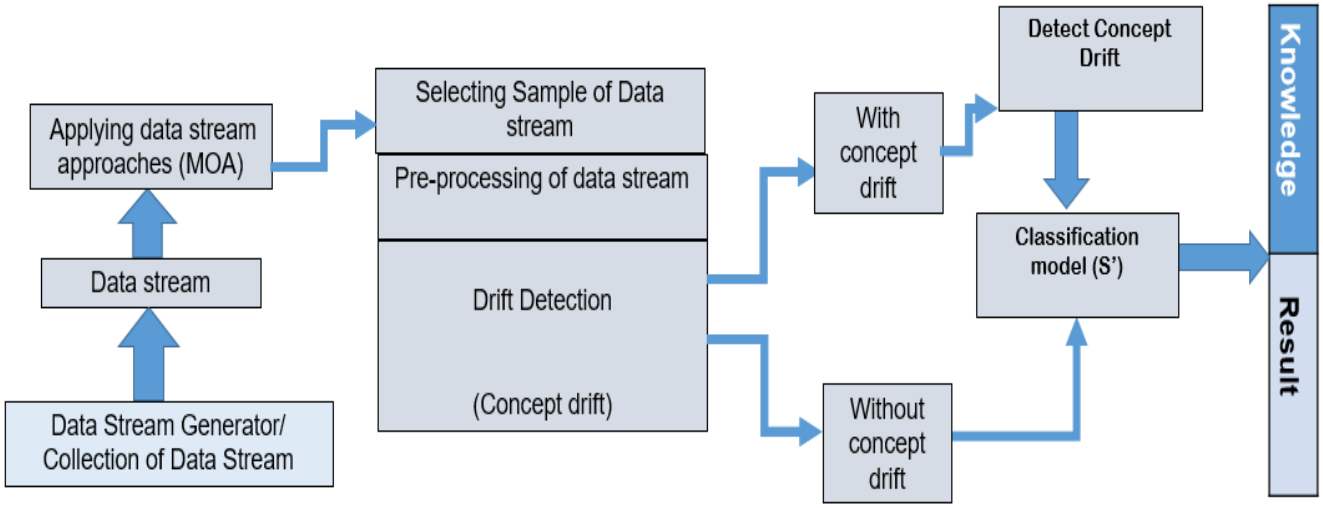

Fig 1: Framework for classification of data stream using with concept drift detection.

3.2 Measures

Accuracy

It is a measure to determine the utility of the dataset [8].

$$\text{Accuracy} = \frac{\text{Correctly classified instances}}{\text{Total number of Instance}} \text{ x } 100 \qquad (2)$$

Kappa Statistics

It measures the agreement of prediction with the true class, formulated as given below:

$$k = \frac{p_0 - p_e}{1 - p_e} \qquad (1)$$

where, p_o is the relative observed agreement, p_e is the hypothetical probability of chance agreement.

4. EXPERIMENTAL DETAILS

4.1 Data Streams

For experimental purpose synthetic and real data streams were taken.

Synthetic data stream: Two synthetic data streams were selected to implement the experiments: LED data stream is composed of 24 categorical attributes. The aim is to predict the digit displayed on a seven-segment LED display, proposed by Breiman et al. in 1984 and Agarwal data stream generated through MOA tool. These data streams are usually used in the concept drift research area [2].

Real data stream: Two real data streams are also taken for experiment: Airline, proposed by Elena Ikonomovska in 2009. It consists 120 million records, containing flight arrival and departure details for all the commercial flights. Poker-Hand is taken from UCI repository and Contain 1, 000, 000 instances.

4.2 MOA (Massive Online Analysis)

This tool was used for classification of data streams and detection of Concept drift [11]

5. RESULTS

5.1 Comparison in terms of accuracy (in Synthetic Data streams)

Table 1. shows the comparison of LED data stream and Agarwal data stream for with concept drift and without concept drift in terms of accuracy with different classifiers. Maximum accuracy with concept drift 73.3% and 90.4% has been achieved using AWE and AUE classifier for LED and Agarwal data streams. For without concept drift 94.7% and 95.1% maximum accuracy has been achieved using Ozaboost and AUE classifiers for LED and Agarwal data streams. Here it was observed that if concept drift is occur in data stream than it will reduce the accuracy. Fig. 2 and 3 show the graphical representation of accuracy for both synthetic data stream.

Table 1. Accuracy of synthetic DS with and without concept drift obtained by different classifiers

Classifiers	LED Data Stream		Agarwal Data Stream	
	With Concept Drift	Without Concept Drift	With Concept Drift	Without Concept Drift
HT	72	92.1	63.4	95.1
NB	47.1	74.8	57.9	88.4
AWE	73.3	80	72.3	93.1
AUE	72.6	94	90.4	95.1
OCBOOST	17	17.2	78.7	93.7
OzaBagAdwin	72.3	73.9	89.6	95.1
OzaBOOST	72.7	94.7	76.2	94
OzaBag	72.3	91.9	66.5	95.1
OzaBagASHT	72.8	73.93	73.9	94.64
HOT	71.8	92.1	72.9	94.3

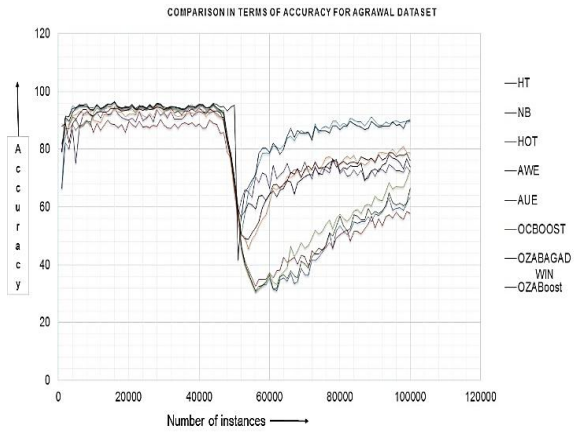

Fig. 2 Comparison of accuracy for LED data stream

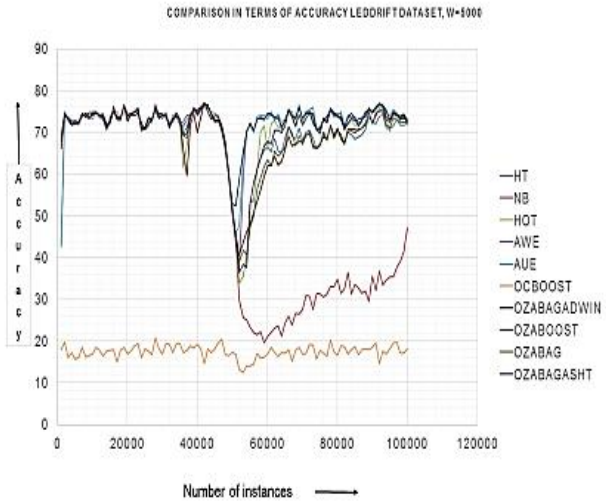

Fig. 3 Comparison of accuracy for Agarwal data stream.

5.2 Comparison in terms of kappa statistics (in Synthetic Data streams)

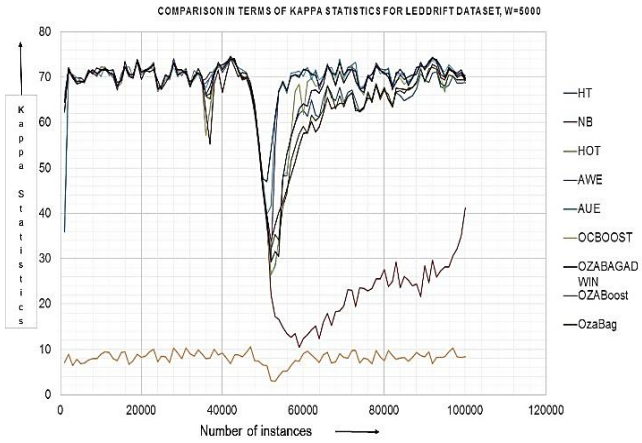

Fig. 4 Comparison of kappa for LED data stream.

Table 2. Shows the comparison of LED data stream and Agarwal data stream with concept drift and without concept drift in terms of kappa statistics with different classifiers. Fig.4 and Fig.5 displays the graphical representation of kappa statistics for both synthetic data streams.

Table 2. Kappa statistics of synthetic DS with and without concept drift obtained by different classifiers

	LED generator		Agarwal generator	
Classifiers	**With concept Drift**	**Without concept drift**	**With concept Drift**	**Without concept drift**
HT	68.84	84.09	20.39	88.81
NB	41.12	71.98	20.55	70.44
AWE	70.29	89.12	40.55	84.34
AUE	69.51	87.92	80.18	88.92
OCBOOST	8	8.15	55.73	85.68
OzaBagAdwin	69.18	70.99	78.49	88.84
OzaBOOST	69.62	89.32	50.22	86.35
OzaBag	69.4	83.69	27.82	86.84
OZABagASHT	69.4	71.04	27.82	87.82
HOT	68.64	84.09	44.73	86.94

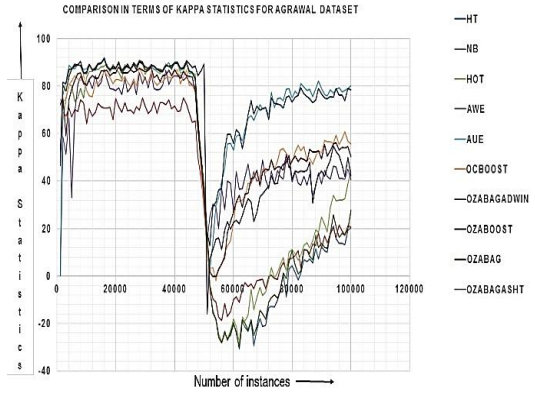

Fig. 5 Comparison of kappa for Agarwal data stream.

5.3 Comparison in terms of Accuracy Time, K.S. (Real Dataset)

Table 3. shows the comparison of airline data stream and poker-hand data stream in terms of accuracy, kappa statistics and time with different classifiers. Fig.5 and Fig.6 show the graphical representation of accuracy and kappa statistics for airline data streams. Sudden drift in the graph shows the occurrence of concept drift in the data streams.

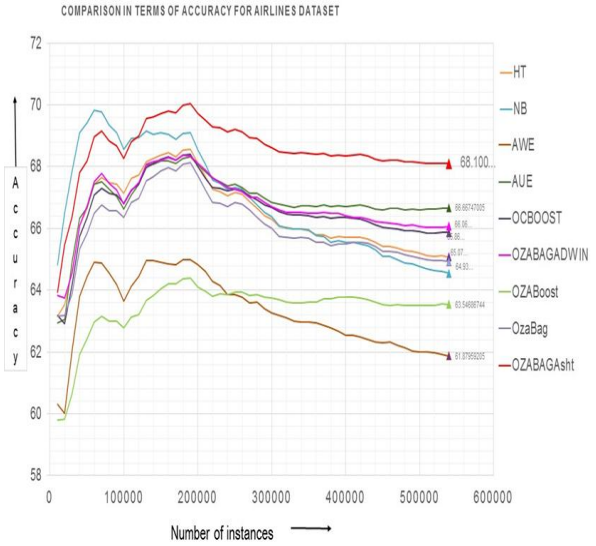

Fig. 5 Comparison of accuracy for airline data stream.

Table 3. Accuracy, Kappa and Time of real DS obtained by different classifiers

	Airline			Poker-Hand		
Classifiers	**Acc.**	**Kappa Stat.**	**Time (sec)**	**Acc.**	**Kappa Stat.**	**Time (sec)**
NB	64.55	25.32	2.2	50	0.0006	4.34
AWE	61.88	21.48	308	28.72	0.09	103
AUE	66.66	31.06	345	62.17	25.95	48.56
OCBOOST	65.86	29.27	56.72	71.15	45.13	73.38
OzaBag-Adwin	66.06	30.2	132.03	85.27	72.34	84.89

OzaBag	60.93	26.17	57.17	85.38	72.56	87.86
OZABag-ASHT	68.1	33.31	49.77	81.87	65.76	58.73
HT	65.08	26.54	4.27	72.14	46.87	6.22
OZABOOST	63.54	25.76	47.43	88.378	78.37	60.50

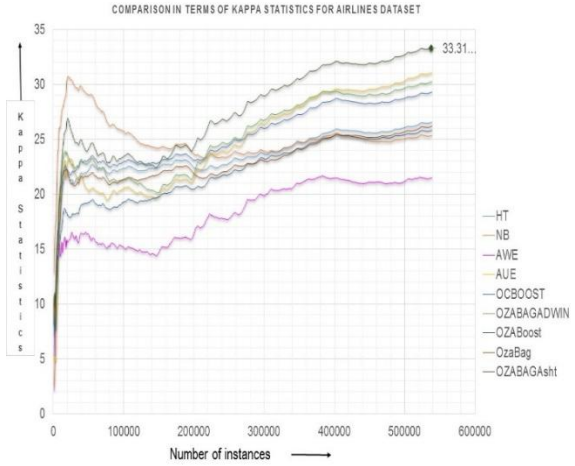

Fig. 6 Comparison of kappa statistics for airline data stream

5.4 Change Detection

In real and synthetic data streams changes (concept drift) are detected by MOA tool and mean prediction error is calculated which shown in Table 4.

Table 4. Concept drift detected in both real and synthetic data streams

Data Stream	Detected changes	Prediction error mean
Airline	253	5.6938
Poker-Hand	8	2.5095
LED generator	18	1.479
Agarwal generator	13	1.21

6. CONCLUSION AND FUTURE SCOPE

The main issue of concept drift in real and synthetic data streams has been investigated in the presented work. When applied to synthetic data stream, in presence of concept drift, maximum Accuracy 73.3% and Maximum Kappa 70.29 has been achieved using AWE classifier in LED data stream. In another synthetic data stream the maximum Accuracy (90.4% with concept drift and 94.70% without concept drift using AUE classifier in Agarwal data stream) and maximum kappa (80.18 with concept drift and 89.32 without concept drift on using Ozaboost classifier in LED data stream) is observed. Similarly, the performance measures has been observed for real data streams. The comparison has revealed that the performance of the classifier depends on the type of data whether it is categorical, numeric or alphanumeric. Further the effects on the privacy preserved classification with concept drift and without concept drift data stream can be compared and analyzed in future.

7. ACKNOWLEDGMENTS

A special thanks is dedicated to all the authors of technical paper for their appreciative original research referred in the initiation of this work. This would have been very difficult to understand and achieve successful completion of the work without the past findings reported by them. This is to put on record the word of appreciation and acknowledgment for the developers of the tools, techniques and thanks for provided easy access.

8. REFERENCES

[1] Kadwe Y. and Suryawansh V., 2015 A Review on Concept Drift, IOSR Journal of Computer Engineering, (JAN-FEB. 2015), 20-26.

[2] Gama J. Zliobait I. Bifet A. and Pechnizkiy M., 2013 A Survey on Concept Drift Adaptation, ACM Computing Surveys, (JAN. 2013).

[3] Mittal V. and Kashyap I., 2015 Online Methods of Learning in Occurrence of Concept Drift, International Journal of Computer Applications, (MAY. 2015), 0975 – 8887.

[4] Bifet A., Read J., Pfahringer B., Holmes G. and Zliobait I., 2013 CD-MOA: Change Detection Framework for Massive Online Analysis, Springer, 92-103.

[5] Hoeglinger S., Pears R. and Koh Y., 2009 CBDT: A concept based approach to data stream, Researchgate, (APRIL 2009).

[6] Bifet A. Read J. , Morales G., and Pfahringer G., 2015 Efficient Online Evaluation of Big Data Stream Classifiers, ACM, (AUG. 2015).

[7] Wankhade K. and Dongre S., 2012 A New Adaptive Ensemble Boosting Classifier for Concept Drifting Stream Data, International Journal of Modeling and Optimization, (AUG 2012), 493-497.

[8] Devasena L., 2014 Efficiency Comparison of Multilayer Perceptron and SMO Classifier for Credit Risk Prediction, Intl J of Advanced Research in Computer and Communication Engineering, (APRIL 2014), 6155-6162.

[9] Cohen E. and Strauss M., 2003 Maintaining time decaying stream aggregates, Proceedings of the 22th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, San Diego, California, U.S.A., (JUNE 2003), 223-233.

[10] Trambadiya T. and Bhanodia P., 2013 A heuristic approach to preserve privacy in stream data with classification, Intl J of Engineering Research and Applications, 1096-1103.

[11] Pramod S. and Vyas P., 2012 Data stream mining: a review on windowing approach, Global Journal of computer science and technology software and data engineering, 27-30.

[12] Chhinkaniwala H., Patel K. and Garg S., 2012 Privacy preserving data stream classification using data perturbation techniques, Intl Conf on Emerging Trends in Electrical, Electronics and Communication Technologies, pp. 1-8.

[13] Li S., Hong L. and Zhen S., 2011 A new classification algorithm for data stream, Intl J Modern Education and Computer Science, 32-39.

[14] Ringne A.G., Sood D. and Toshniwal D. 2011 Compression and privacy preservation of data streams using moments, Intl J of machine learning and computing, 473-478.

[15] Benjamin M. Fung , Wang K. , and Philip S., 2007 Anonymizing classification data for privacy preservation, IEEE Trans on Knowledge And Data Engineering, 711-725.

[16] Aggarwal C. and Philip Y., 2008 A general survey of privacy-preserving data mining models and algorithms, Springer, 11-52.

[17] Street W. and Kim Y., 2001 A streaming ensemble algorithm (SEA) for large-scale classification, In KDD 01 New York, NY, USA, ACM Press, 377-382.

[18] Babcock B., Babu S., Datar M., Motwani R. and Widom J., 2002 Models and Issues in Data Stream Systems, ACM PODS Conference.

[19] http://moa.cms.waikato.ac.nz/