IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-10, Oct 2018 E-ISSN: 2347-2693

Empirical Analysis on Stream Classification & Clustering with Concept Drift in MOA

Hari A. Patel[1], Harsh N. Patel[2], Nirav Bhatt[3*]

[1,2,3]Dept. of Information Technology, CSPIT, Charotar University of Science and Technology, Changa, Gujarat, India

[*]*Corresponding Author: niravbhatt.it@charusat.ac.in, Tel.: +91-9998-582812*

Available online at: www.ijcseonline.org

Accepted:18/Oct/2018, Published:31/Oct/2018

Abstract— Stream data processing is the next 'big thing' in big data which is one of the most propagating fields in computer science. The stream data analytics is an important aspect while dealing with data stream mining. While dealing with the classification of stream data, concept drift and its effect are required to be considered. Massive online analysis (MOA) is one of the most popular tools to perform analytics on stream data. We primarily deal with three features which are provided by moa namely classification, clustering & concept drift. The key emphasis is on experimental analysis on the combination of different procedures and learner algorithm which are suited for training the model so it can be used for the prediction purpose. Besides that, we have also tried to identify drift (change) in data and its effect on performance. So conceptually after taking proper measures about the noise and drifts, we can construct a model which is persistent to all the changes it faces. Stream analytics also required exploring the different clustering techniques which have a wide number of applications. We have presented all the empirical analysis carried out on classification and clustering techniques in a tool called MOA.

Keywords— Stream Processing, Concept Drift, classification, Clustering, MOA.

I. INTRODUCTION

Massive Online Analysis (MOA) is a software environment for implementing algorithms and running experiments for online learning from evolving data streams [1].MOA (Massive online analysis) can be said as an open-source framework software which allows to build and run experiments of machine learning or data mining on evolving data streams . The Classification techniques commonly build models that are used to predict future data trends [11].

MOA consists of different learner algorithms and also different stream generators which can be encountered from the graphical user interface or command line version. MOA is built on experience with both WEKA and VFML [3]. we deal with three features of MOA Tool namely Classification, Concept Drift, and Clustering. The in-depth descriptions of all the features are described in the following section. Data mining also called information mining or certainly finding is the term which is utilized for removing or finding helpful data from the information that is available in vast databases [22]. Following section includes stream processing with its algorithms which is followed by comparison of different classifiers and experiments of concept drift.

II. STREAM PROCESSING

Stream processing is a kind of technology that will allow a user to do processing or the execution of a query on the continuous data stream and this will help in detection or the prediction of the conditions in a small span of time after the data is received. It does not provide any option for storage; the incoming data is directly processed as soon as it arrives. The time period for this may range from milliseconds to seconds. Let us consider an example where we want to perform query or operations on the data stream which is arriving from the satellites for the weather prediction, so we can develop a model with help of tool which can predict the weather from the testing on our trained model. This Stream processing is also called real-time streaming analytics, complex event processing.

III. LEARNER ALGORITHM

Table 1: Algorithms description

Algorithm Name	Description
Naïve Bayes	When one has hundreds of thousands of data points and few variables in one's training dataset. If time is the main factor for one's situation then Naive Bayes is one of the fastest classification algorithms. It works on the principles of Bayes theorem to predict the class of unidentified dataset [4]. It is a classification technique based on the principles of Bayes Theorem with an assumption of independence among predictors [4]. In simple language, the Naive Bayes Classifier assumes that the presence of one feature is independent of any other features of the class. The principal disadvantage of this algorithm is that can't support Concept drift, the algorithm is not retrained over the time [15].
Hoeffding Tree	Hoeffding trees were introduced by Domingos and Hulten in the paper "Mining High-Speed Data Streams" [10]. The Hoeffding tree induction algorithm implies a decision tree from a data stream incrementally, briefly checking each

	example in the stream only one time, without the necessity to store examples after it has been used to update the tree. Hoeffding Tree can often be enough to choose an optimal splitting attribute by Hoeffding Bound [20]. The only piece of information required in memory is the tree itself, which stores enough information in its leaves in order to grow, and can be used to form predictions at any time between processing training examples [21].

IV. EVALUATION PROCEDURES

Table 2: Procedures description

Procedure Name	Description
Holdout	Holdout method is of great use when the partition process between training and testing sets have been predefined. Now for the evaluation process of the performance of the respected model, we can evaluate it systematically i.e after a specified amount of training examples. Consulting about examples of the holdout, we can take into consideration the stream such that it has not been used to train the learner algorithm. In this, a procedure can look ahead to collect a batch of examples from the stream for the use as test examples, and if efficient use of examples is desired they can then be given to the algorithm for additional training after testing is complete [21].
Interleaved Test-Then-Train	This is a way for the evaluation and it is to interleave testing with training. In this type of procedure, every concerned particular will be used for testing the model before it goes for training and so we can see that the accuracy will be incrementally updated. When we are using this method the concerned model will be tested on the examples it is unaware of. The example of this procedure is no holdout set is needed for testing; it makes the use of currently available data and makes maximum use of it. Here we will get a plane plot of accuracy over the factor time. But this procedure has a disadvantage of distinguishing and measuring training and testing time.

V. CONCEPT DRIFT

Concept drift means when a sudden change in the dataset occurs which was not predictable. A majority of concept drift research in data streams mining is done using traditional data mining frameworks such as WEKA [9]. One of its examples is weather prediction. The distribution generating the items of a data stream can change over time. To summarize, three basic approaches to handling concept drift can be distinguished: instance selection, instance weighting, and ensemble learning [6]. These changes, depending on the research area, are referred to as temporal evolution, covariate shift, non-stationary, or concept drift. The general assumption in the concept drift setting is that the change happens unexpectedly and is unpredictable [12].

There are mainly three type drifts namely Original drift, real drift and virtual drift.

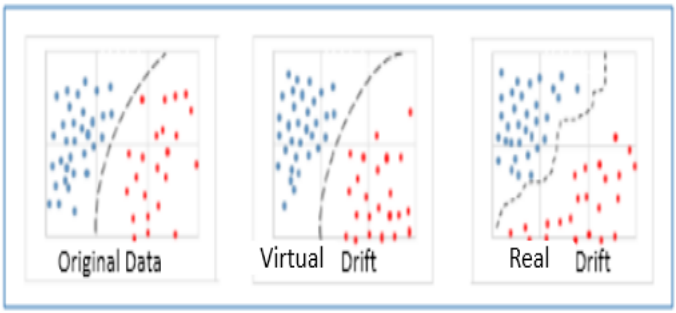

Fig – 1 Real Vs. Virtual Drift

There are also some other types of drift namely Sudden, Incremental, and Gradual, Reoccurring, Blip, Noise. For some Cases, misclassified streams produce noise in data and due to the noise, the drift will occur [13].

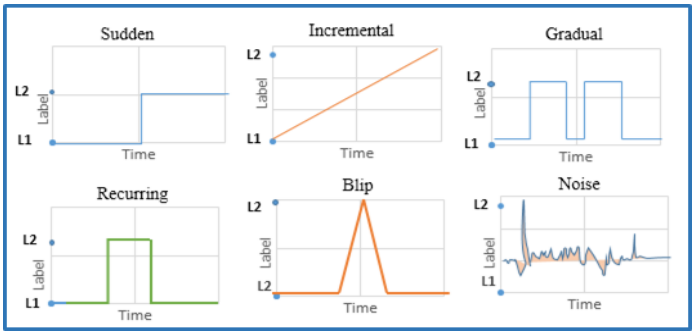

Fig – 2 Different Types of Drift

VI. COMPARATIVE ANALYSIS (CLASSIFICATION)

Dataset Description:

1. Forest Covertype: It contains 581, 012 instances and 54 attributes. Here Covertype is the target attribute (means this the attribute whose value will be predicted from the given 54 attributes).

2. Poker-Hand: In this dataset, there are 1,000,000 numbers of instances and 11 attributes. There is one class attribute that describes the "Poker Hand" and that is our target attributes.

3. Electricity: The ELEC dataset contains 45, 312 instances. The class label identifies the change of the price relative to a moving average of the last 24 hours.

4. Airlines Dataset: The task is to predict whether a given flight will be delayed, given the information of the scheduled departure i.e the target variable is Arrival Delay (in seconds).

Now, suppose the evaluation procedure is Interleaved Test Then Train, learner algorithms are the tree, HoeffdingTree, Naïve Bayes, and Ruler Classifier on performing the classification test on the datasets we get the following table which shows accuracy and time.

Table 2 - Comparative Analysis with InterleavedTestThenTrain

InterleavedTestThenTrain									
	NaiveBayes			HoeffdingTree			RuleClassifier		
	Poker	Electricity	CovTypeNorm	Poker	Electricity	CovTypeNorm	Poker	Electricity	CovTypeNorm
Accuracy	59.55	**73.26**	60.52	76.07	79.2	**80.31**	62.95	**73.21**	67.45
Time (in seconds.)	7.2	**0.23**	14.95	9.66	**0.48**	19.67	10.41	**6.62**	8.36

Here, Electricity dataset has the lowest numbers of instances and integer type which is favorable for Naive Bayes thus it gives the best result. Hoeffding Tree gives the best accuracy when there is a large number of attributes and instances thus it performed best for covertypeNorm.Rule classifier works on the conditional classification which works best with electricity dataset thus it gives good accuracy and less amount of time.

Now, suppose the evaluation procedure is Prequential, learner algorithms are the tree. HoeffdingTree, Naïve Bayes, and Ruler Classifier for performing the classification test on the datasets we get the following table which shows accuracy and time.

Table 3 - Comparative Analysis with Prequential

Prequential									
	NaiveBayes			HoeffdingTree			RuleClassifier		
	Poker	Electricity	CovTypeNorm	Poker	Electricity	CovTypeNorm	Poker	Electricity	CovTypeNorm
Accuracy	39.1	75.3	**81.5**	82.5	81.6	**87**	36.6	73.7	**80.8**
Time (in seconds.)	0.47	**0.16**	1.84	0.45	**0.28**	2.11	11.7	**6.62**	8.72

Prequential method for classification has the concept of sliding window which works best with high no instances and attributes so covtypeNorm dataset has high accuracy across all algorithm.

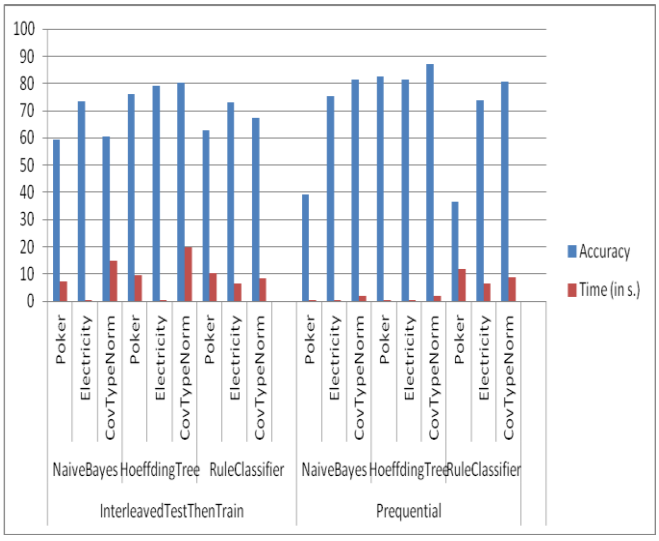

Fig – 3 Comparative Analysis Graph

So here we can observe that we get higher accuracy in the case of trees. HoeffdingTree learner algorithm and InterleavedThenTrain method.

Noise is also one type of drift thus it is a problem of classification and we have to train our model with drift so in letter stages it has good accuracy. Here, we have tested two learner algorithms with generator called SEA generator with 0 and 60 percent noise. The Output is as follows:-

Table 4- Experimental Data (Concept Drift)

	Naïve Bayes	Hoeffding tree	Naïve Bayes (Noise 60%)	Hoeffding tree (Noise 60%)
Accuracy	94.55	98.62	59.76	59.18
Time (in s.)	0.17	0.30	0.19	0.48

VII. CLUSTERING

Clustering means the method by which we can split the set of the desired data or also elements into meaningful similar subclasses called clusters. We can say clustering is the grouping process of the set of objects in a proper way such that the entities in the equivalent group are attached and located together. It is also called cluster analysis. Cluster analysis is a fundamentally exploratory tool that seeks to sort data vectors into like groups when true group memberships are not known [19]. If one wants to evaluate the goodness of a given clustering algorithm, the corresponding approach, i.e. its underlying clustering model, is of importance [14].

The data stream clustering problem is defined as to maintain a continuously consistent good clustering of the sequence observed so far, using a small amount of memory and time [5].

The application of clustering algorithms to data streams has been concerned with either object-based clustering or attribute-based clustering, with the former being far more common [18].

Its applications include World Wide Web, Image Analysis, Recommendation Systems, Fraud Detection, and Marketing. Most available static data are becoming more and more high dimensional. Therefore, subspace clustering, which aims at finding clusters not only within the full dimension but also within subgroups of dimensions, has gained a significant importance [7].

Our goal is to build an experimental stream clustering system able to evaluate state-of-the-art methods both regarding clustering algorithms and evaluation measures [2].

In real time data analytics, these stream clustering is of great use. Now for the testing purpose of the obtained clustering, various measures and parameters are considered which are provided for checking the quality of the data which will be reflected by the cluster, all these features are provided in the MOA tool. In MOA there is a special provision provided in a separate tab where you can compare different clustering algorithms and also get the information about its different measures that too in both the ways graphical and numerical.

VIII. STREAM CLUSTERING ALGORITHMS

Table 5- Clustering Algorithms description

Algorithm Name	Description
CluStream	This method uses micro clusters to maintain the statistical information about the data. The micro-clusters can be considered as the time-dependent addition of cluster feature vectors [8]. For the storage of these micro-clusters, it is done by using snapshots in time and this follows the pyramidal pattern. With the help of this pattern summary statistics from different time can be evaluated. An effective technique, easily distributable over various computing nodes [16].
StreamKM++	In this method, a small weighted sample of the data is used and with the help of the k-means++ algorithm as a randomized seeding technique, the first values of the cluster are chosen. The main drawback of such k-means based data stream clustering algorithm (Clustream) is that user has to give no. of the cluster (k) in advance [17].

IX. CLUSTERING DEMO (STREAM GENERATOR)

Now we will see how clustering with stream generator that generates a random radial basis function stream.
We can see in detail how by selecting different forms the clusters will be displayed.

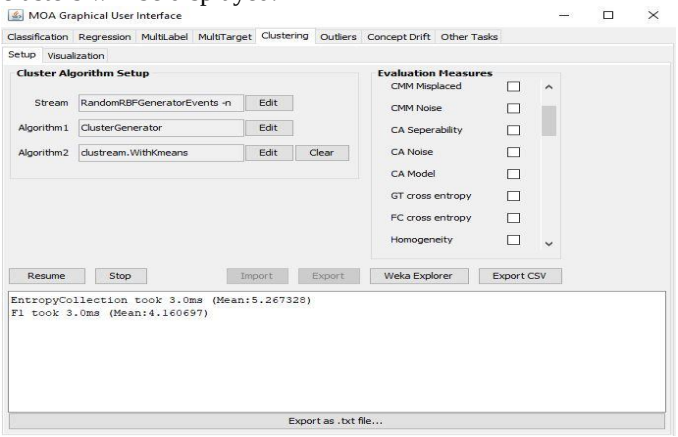

Fig – 4 Clustering Setup window

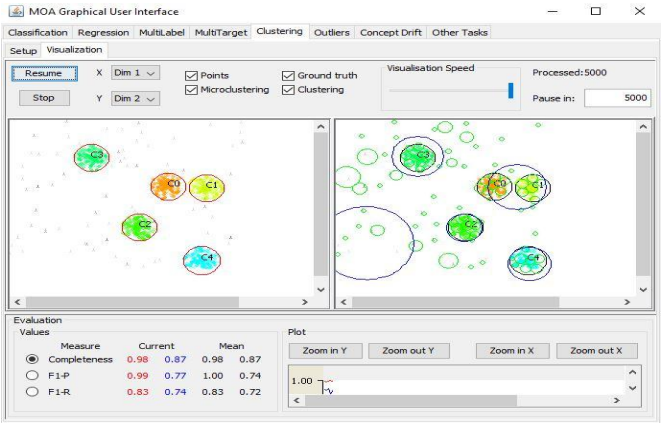

Fig – 5 Clustering Output window

So here as it can be seen three different measures are observed with their respective values, namely completeness, F1-P, F1-R and the values for both the algorithms are seen. Red colored values are for the cluster generator algorithm and Blue colored are for the cluster with k-mean algorithm.

X. CONCLUSION

With the help of MOA (Massive Online Analysis) tool experimental analysis was performed for the classification purpose and as an outcome, we got the learner algorithms and procedures which can

be best for the purpose of training our model such that it can predict the future circumstances and can be used for real-time data analytics. After this, with the help of concept drift topic we got an overview about how the noise can affect our prediction model and hence we trained our model in such a way that though there is a sudden change in the data, our model can survive the changes. At last with help of clustering, we studied cluster analysis which helps in different aspects of the real world. We conducted different experiments on ARFF file and random event generator to observe how clusters are formed using different clustering algorithms.

REFERENCES

[1]. Bifet, A., Holmes, G., Pfahringer, B., Read, J., Kranen, P., Kremer, H., ... & Seidl, T. (2011, September). MOA: a real-time analytics open source framework. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 617-620). Springer, Berlin, Heidelberg.

[2]. Kranen, P., Kremer, H., Jansen, T., Seidl, T., Bifet, A., Holmes, G., & Pfahringer, B. (2010, December). Clustering performance on evolving data streams: Assessing algorithms and evaluation measures within MOA. In Data Mining Workshops (ICDMW), 2010 IEEE International Conference on (pp. 1400-1403). IEEE.

[3]. Kranen, P., Kremer, H., Jansen, T., Seidl, T., Bifet, A., Holmes, G., ... & Read, J. (2012, April). Stream data mining using the MOA framework. In International Conference on Database Systems for Advanced Applications (pp. 309-313). Springer, Berlin, Heidelberg.

[4]. Jani, R., Bhatt, N., & Shah, C. (2017, March). A Survey on Issues of Data Stream Mining in Classification. In International Conference on Information and Communication Technology for Intelligent Systems (pp. 137-143). Springer, Cham.

[5]. Gama, J. (2010). Knowledge discovery from data streams. Chapman and Hall/CRC.

[6]. Tsymbal, A. (2004). The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin, 106(2).

[7]. Hassani, M., Kim, Y., & Seidl, T. (2013, April). Subspace MOA: subspace stream clustering evaluation using the MOA framework. In International Conference on Database Systems for Advanced Applications (pp. 446-449). Springer, Berlin, Heidelberg.

[8]. Ahmed, M. (2018). Data summarization: a survey. Knowledge and Information Systems, 1-25.

[9]. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. SIGKDD Explor. Newsl., 11(1):10–18, 2009.

[10]. Domingos, P., & Hulten, G. (2000, August). Mining high-speed data streams. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 71-80). ACM.

[11]. Conference on Knowledge Discovery and Data Mining, pages 71–80, 2000.

[12]. Al-Radaideh, Q. A., & Al Nagi, E. (2012). Using data mining techniques to build a classification model for predicting employees performance. International Journal of Advanced Computer Science and Applications, 3(2).

[13]. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM computing surveys (CSUR), 46(4), 44.

[14]. Hoens, T. R., Polikar, R., & Chawla, N. V. (2012). Learning from streaming data with concept drift and imbalance: an overview. Progress in Artificial Intelligence, 1(1), 89-101.

[15]. Kremer, H., Kranen, P., Jansen, T., Seidl, T., Bifet, A., Holmes, G., & Pfahringer, B. (2011, August). An effective evaluation measure for clustering on evolving data streams. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 868-876). ACM.

[16]. Ibanez, A. C. (2017). Introduction to Stream Mining.

[17]. Devi, Y. S., & Nagababu, G. (2017). Comparison of Clustering Algorithms in Data stream mining: A Literature Survey.

[18]. Dipti, M., & Patel, T. (2014). K-means based data stream clustering algorithm extended with no. of cluster estimation method. International Journal of Advance Engineering and Research Development (IJAERD), 1(6).

[19]. Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., De Carvalho, A. C., & Gama, J. (2013). Data stream clustering: A survey. ACM Computing Surveys (CSUR), 46(1), 13.

[20]. Wilks, D. S. (2011). Cluster analysis. In International geophysics (Vol. 100, pp. 603-616). Academic press.

[21]. Parikh, D., & Tirkha, P. (2013). Data mining & data stream mining—open source tools. International Journal of Innovative Research in Science, Engineering and Technology, 2(10), 5234-5239.

[22]. Fernandes, M. (2017). Data Mining: A Comparative Study of its Various Techniques and its Process. International Journal of Scientific Research in Computer Science and Engineering, 5(1), 19-23.

Authors Profile

Hari A Patel pursing Bachelor of Technology from CHARUSAT University in the stream of Information Technology. His reasearch area includes big data analytics and stream data analytics.

Harsh N Patel pursing Bachelor of Technology from CHARUSAT University in the stream of Information Technology. His reasearch area includes big data analytics and data mining.

Nirav Bhatt is working at Department of Information Technology in Chandubhai S Patel Institute of Technology, CHARUSAT. He had received degree of Master of Engineering in Computer Engineering from Dharmsinh Desai Institute of Technology and currently pursuing his Ph.D. in the area of Big Data Stream Analytics. His research interests include Database System, Data Mining and Big Data Stream Analytics. He is also a member of Computer Society of India and ACM and member of ACM Chapter at the institute. He is also coordinator of SWAYAM-NPTEL Local Chapter which is the National MOOCs portal being developed by MHRD, Govt. of India.