ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse185932020.pdf
https://doi.org/10.30534/ijatcse/2020/185932020

Impact of Variable Size Chunk Data on Classifier Performance using Ensemble Techniques

[1]Leena Deshpande, [2]M R Narasingarao
[1]Research scholar, KLEF Deemed to be University
[1]Department of Computer Science and Engineering, Vijayawada, India
leena.deshpande@viit.ac.in
[2]Professor, KLEF Deemed to be University
[2] Department of Computer Science and Engineering,
KLEF Deemed to be University , Vijayawada, India
ramanarasingarao@kluniversity.in

ABSTRACT

In recent times, enormous growth of real-world data and its usage raised an issue of processing data for extracting meaningful patterns. Due to huge volumes and diversification in the data, traditional knowledge of data mining algorithms lower down the accuracy. Due to data drift, a good accuracy model may not outperform for generalized data. So, it is necessary to handle the causes of drift and its impact on the model accuracy. Existing approaches use a fixed size sliding window approach. In contrast, our approach uses both fixed window and adaptive window approach to detect the concept drift. We have used maximum likelihood estimation technique. CUSUM chart ,Simple Moving Average and a cross correlation technique to detect a change in the concept . We have analyzed the impact of variable size chunk data on different ensemble model. Our approach improves the classifier accuracy using better feature selection and evolution method. The combine approach of variable sized sample and weighted Ensemble classifiers not only detect the change in the concept but also applied for drift detection. Under data drift strategy, we inclusively compare the classifiers performance on electricity dataset ,refereed by research community .

Key words : Accuracy, Concept Drift, Data Drift, Data Mining, Ensemble -Classification

1. INTRODUCTION

The massive and varied data is generated over the network due to wide usage of device communication channel. This data is generated in various formats like audio, video, text which flow from one end to another as a data stream. Since the data is generated from heterogeneous sources it becomes dynamic in nature. Due to such dynamic nature of data, the traditional tools and techniques do not fit for such varied data to accurately predicts the unknown pattern. The performance and efficiency of these data raises the new challenges in data mining community. Supervised technique is one of the most significant tasks in data mining. Machine learning lexicon based and linguistic analysis methods have been suggested by research community. In case of large volume of data and varied patterns, decision boundary of the classifier is become more complex. In such situations, a more generalized solution is required. In continuous streaming data which we call as non stationary data causes arbitrary changes in the patterns are seen in big data applications like weather forecasting, fault detection, intrusion detection system, chemical reactor plants etc . Also the applications where users comment and share their views on social media like Twitter Facebook causes drastic change in the behaviour and pattern. Such pattern identification techniques refer to rigorous training of streaming data, it also needs attention to the optimum use of computational storage and memory usage. Due to data drift, traditional knowledge of machine learning models lower down the classifier's accuracy. Broadly a classification is divided into two techniques, a single model and an ensemble classification Ensemble learning approach like Bagging, Boosting, Adaboost are widely used and proves more better over the single expert system techniques.[1]. In non-stationary environment, finding good set of features from the underlying data is an important task for identifying drift. Existence of drift occurs due to change in distribution of data which further degrades the classifier accuracy. Gama et.al [2] has broadly classified the concept drift into 3 categories. i)Sequential based ii) Statistical based and iii)Window based . In this paper based on above methods,

We have proposed new concepts for sample selection, diversification in the data, and impact of variable size chunk in classification process. Diversity in data leads to a drift. Several classification algorithms have been evolved to deal with sudden, gradual or recurring drift in data streams. In this paper, we describe and identified data drift and analyses its effect on classifier. The huge data distributed over large scale needs to distribute using clusters. The goal of clustering is not only to group the data but also to reduce high dimensional data to discover the meaningful information from it. In this paper, we performed a systematic analysis on 2 different data sets . we have used a simple but appropriate cross correlation methods and CUSUM method for selecting the variable size

chunk of data. A predefined threshold represents the change in the concept if observed values are deviating from the threshold value. The combine approach of variable sized sample and weighted Ensemble are used for classification. These methods are applied on both numeric darta. The paper is divided into 4 sections System architecture is explained below. The remaining section of the paper is configured as Related work, Methodology, followed by performance analysis.

1.1 System Architecture

The proposed system architecture explained in Figure 1 is based on divide-and-conquer approach. It is mainly divided into 2 architecture components. In the first component, various chunk of data has been extracted based on important feature referred as clustering. Another component is a supervised learning method consisting of the training data containing a set of an input-output pair, for analyzing new unlabeled data. An input data is partitioned into three chunks using clustering techniques, named as: chunk 1, chunk 2 and chunk 3. Each cluster or chunk will have the most important feature based on data. Features are extracted based on cross correlation and threshold difference with the previous chunk of data . The CUSUM chart will be used to assign weights to the stream of data based on observed variation in the data stream. Finally, the class label of new test data will be predicted with the help of local voting mechanism performed on each chunk.

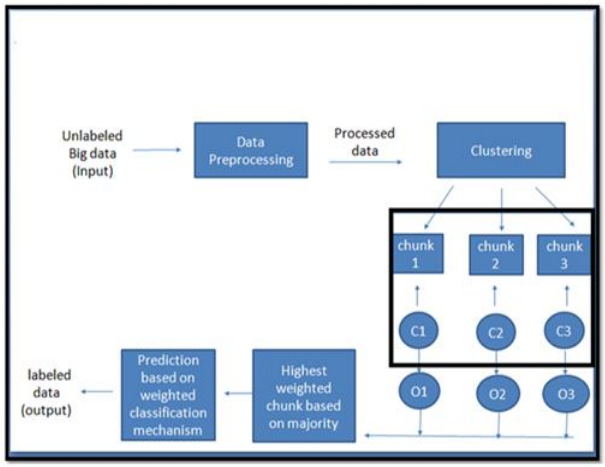

Figure 1: System Architecture

2. RELATED WORK

Existing research in concept drift classification focuses on concept drift causes and efficiency of the classification model. Broadly a classification is divided into two techniques a single model and an ensemble classification. Data drift is found due to diversified and unlabeled data in the stream and hence misclassify class labels. Several classification algorithms have been evolved to deal with sudden, gradual or recurring drift in data streams. To address the unlabeled infinite length data, a single model-based technique is used to identify the most recent concept and to locate the drift. Researchers drift detection work is broadly divided into three categories, sampling the stream of data, measure the properties of the samples data and classify the data using classification. For sampling, Cesario et. al [1] in their work, proposed the bootstraps aggregating techniques in which original training data is reduced to generate bootstrap samples An Ensemble technique proposed by Street N and Kim Y[3] uses clustering technique where multiple clusters are generates and of same size to earn improved results than individual clusters. A voting function is applied to retrieve novel sample. A vote based system in Jie Hu et al[4]. and Liping Jing et al.[8] is used for class prediction where a class with ultimate count of votes is served as final prediction. Laradji et al.[5] proposed Majority Voting Ensemble (MVE) technique. In this technique each class receives a vote based on class prediction and final count of each vote will serve the best accepted label of a class. B. Parker et al. have proposed a HSMiner algorithm which uses a hierarchical approach and weighted voting and boosting method to improve the individual classification accuracy. A composite Ensemble model is built using eight independent classifiers. Similar approach proposes by Peipei Li et al.[6] where an optimal clustering algorithm is applied on dataset and further given to classifier for optimum solution using Ensemble clustering method. A Hierarchical Voting Classification Scheme is proposed by Littlestone and Warmuth.[7] A final prediction is the estimation of individual base class accuracy. All of the these approaches uses clustering alone as a main element but unfortunately clusters alone are not capable of classification process as they contribute more towards categorizing the samples, reducing the high dimensional features and pre-processed the samples . However, the data does not give absolute class label knowledge as large number of unsampled data still exist. For producing the correct label output, few have given attention to automatically generate class label. A dynamic weight based and ensemble based approach is used for labeling and classifying the data. Zhang et. Al[7] proposed an approach of segregating huge data using small clusters Then the individual base classifiers are generated to train these clusters locally. in their further work[7], authors proposed an active learning framework which uses a minimum variance method for instant labelling of data.

For drift detection, commonly used technique is windowing technique. Drift is also measured by detecting outlier in a stream of data . A set of outlier can form a novelty in the underlying data. In novelty detection, the dynamic weight updating strategy and retraining the old model with new one ,maintain high generalization of old and new concepts and provides high accuracy Masud *et al.*[9] detect the concept evolution problem by proposing adaptive threshold for detecting outliers .Oza *et al.* [10]proposed weighted majority algorithm. Weights are updated for wrong

inputs and recursively applied to improve the accuracy. A k-NN based classifier is proposed by Choensawat *et al.* [11] for identifying novelty class and outliers in real estate project. Snehalata et al [12] proposed a hybrid approach known as HEDDM which reacts to both type of drift, sudden and gradual drift.

Social media platforms gives rise to big data formation and different context. For better sentiment and context analysis, a major challenge lies in data preprocessing of such huge data. A detection of spatial trajectories of moving objects can be discovered using Social networking sites data in our case a novelty is detected after pre-processing and classifying tweets. e.g. in the paper by Senaratne *et al*. [13], authors proposed work on tweet data based on famous singer Lady Gaga's tweet. Spatiotemporal data comes with large amount of complex and hidden knowledge. These data become a point of focus for classification problem. spatial trajectories of her world tour are detected by tweets. For detecting her twitter activities, a framework of kernel density estimation for detecting hotpot clusters of twitter has been used . Sing T and kumara M [14], analyses insignificance and significance of slang word used in 2014 election data , held in India. Author detects a concept drift using sliding window protocol for finding trending topics to predict election results. Zhao *et al.* put forward the accuracy of recommendation by addressing social influence of users. A user interaction and time distance is used as a weighing factor. A matrix factorization technique is used as a temporal dynamic for identifying social relationships. Most of the literature survey is found on implementing a single classifier and not on ensemble. Our work used numeric data for drift detection using ensemble. A statistical based approach for detecting drift is an alternative method which is discussed by research community. Researchers contribute not only on high dimensional data but also the computation and resources used to handle high dimensional data. Muhammad et al[15] proposed a Monte Carlo tree search (MTCS) method for retrieving features from very high dimensional data. Multiple change detection methods have been proposed for categorical data stream by Joshua Plasse and Niali Adams [16]. Kyosuke Nishida and Koichiro Yamauchi [17] developed a STEPD algorithm which uses a statistical test for equal proportion for online environment. However this test is not suitable for small and imbalanced data . Danilo Rafael et al.[18] proposed a Fishers Exact for detecting gradual and abrupt drift. The test is computationally costly over STEPD algorithm. Pavlos S. Efraimidis [18] used a weighted random sampling A-chao and A-ES algorithms were used. KL divergence[19] is another non symmetric method used for detecting deviation from the probability distribution. The high dimensional diversified data is treated with GA-Adaboost feature selection scheme with the use of Ensemble learning [20] which maximizes the performance . Similarly a novel based feature selection approach known as aspect based classification is proposed by Maganti and Nalini [21]. The method proposes the aspect of positive and negative opinion on diverse domain.

3. METHODOLOGY

A. Algorithm

Input: Training chunk S_i ,Test chunk S_{i+1} and S_{i-1} ,threshold T, Ensemble of classifiers Z_i-1 built from $t_i = 0$ to $i - 1$ AP ,set of all positive instances accumulated from the previous training chunks from $t_i = 0$ to $i - 1$

Output: Ensemble Zi , for the test chunk S_{i+1}

1. Begin for each timestamp $t_i = 1,2....$ do
2. Split the current training chunk Si into samples having high cross co-relation
3. if corel(S_i) > Theshold(T) then then TrainSet = Bi ∪ AP
   else
   Find next chunk nearest to threshold T.
4. Estimate weights of each classifier
5. Calculate weighted average of all the n base models;
6. Build the ensemble model E

Data Sets

We carried experiments on benchmark dataset .We have used numeric data of electricity. The experiment is carried using Windows 7, intel i5 processor

Electricity data :-This data sets is publicly available on MOA (Massive online Access) framework. [22] It is widely used as a popular benchmark dataset collected from Australian New South Wales Electricity Market containing 45312 instances dated from May 1996 to December 1998 .Each example of the dataset refers to a period of 30 minutes It is a binary classifier which predicts the price of electricity as Price UP and Price Down. Due to uncertainty in consumption of electricity ,prediction model can decide the features responsible for sudden rise or fall in the consumption. Each example on the dataset has 5 fields, the day of week, the time stamp, the NSW electricity demand, the Vic electricity demand, the scheduled electricity transfer between states and the class label, which analyse the variation in price values in terms of class label as UP or DOWN by using moving average of last 24 hours The class label(UP and DOWN) identifies the change of the price relative to a moving average of the last 24 hours. In this datasets the peak utilization of electricity is from 30 to 40 timeslot.

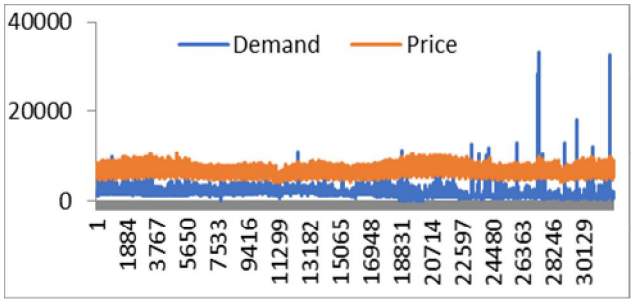

Figure 2: Fluctuation in average demand

Below graph shows fluctuation in average demand. Though demand is flexible in few slots, price remains stable for all the slots. Blue line indicates the average demand and orange represents the price.

Our work in this proposed model is to identify which attributes gives the insight for the fluctuation. We have categorized no of samples and applied statistical test as shown below. Following table (Table 1 and Table 2) summarises the key attributes and classification-based approach with applied threshold value. Samples decide the chunk size and relevant attribute selection for electricity dataset .An increase in the sample size, parameter selection, approximation based approach leads to a better accuracy in terms of prediction. In our work the proposed model predicts average accuracy when we consider all tuples including repetitive value and hence to improve it further, a cross correlation is used for Price & Time slot, Price and Demand.

*De*mand and Time Slot. A simple moving average and exponential moving average [23]is applied on 4 parameters, Day, Time Slot, Price, and Demand. Based on these parameters the accuracy of individual and ensemble is calculated. Out of 45312 instances, a sample size 250 to 750 is selected as 250 with increment of 100 samples so the samples between 250 to 750 will be 250,350,450, 650 and 750 respectively. We have applied Principal Component Analysis for feature selection. Attributes having lower variance are taken and given to the model for training.

Table 1: classifier accuracy on overlapping data

Chunk Size(N)	Training Samples	Window Size	Naive Bayes	KNN	Logistic
N=250	45312	Fixed	72.61	70.81	75.45
N=250	10384	Fixed	54.6	62.14	78.37
N-250	10372[	Fixed	72.61	70.81	75.45
N=500	45312	Fixed	72.61	79.09	75.45
N=500	10384	Fixed	54.6	62.14	78.37
N=500	10384	Fixed	72.7	68.03	75.54
N=500	10372	Fixed	72.43	79.09	76.14
N=250 to 750	45312	Variable Size	72.43	79.09	76.14
N=250 to 500	10372	Variable Size	72.43	79.09	76.14
N=250 to 750	10372	Variable Size	72.43	79.09	76.14
N=250 to 500	10384	Variable Size	72.43	79.09	76.14
N=1000	10384	Fixed	72.7	68.03	75.54
N=1000	45312	Fixed	72.43	79.05	76.14
N=1000	10372	Fixed	72.7	68.03	75.54

Table 2: Ensemble Classifier accuracy on overlapping data

Chunk Size(N)	Training Samples	Window Size	Ensemble
N=250	45312	Fixed	**76.08**
N=250	10384	Fixed	**89.84**
N-250	10372[	Fixed	**79.44**
N=500	45312	Fixed	**76.07**
N=500	10384	Fixed	**89.84**
N=500	10384	Fixed	**79.58**
N=500	10372	Fixed	**78.42**
N=250 to 750	45312	Variable Size	**78.01**
N=250 to 500	10372	Variable Size	**89.47**
N=250 to 750	10372	Variable Size	**91.58**
N=250 to 500	10384	Variable Size	**78.42**
N=1000	10384	Fixed	**79.58**
N=1000	45312	Fixed	**78.01**
N=1000	10372	Fixed	**79.58**

Table 2 shows the Ensemble classifier accuracy with overlapping data . Ensemble proves better over KNN ,Naïve Bayes and Logistic .

We have used CUSUM chart which is based on maximum likelihood estimation given as $Sm = \Sigma_{i=1}^{m} (xi - \mu 0)$ **(1)** where Sm is a cumulative score for xi and µ0 is the target mean . Though the process is under statistical control it shows the deviation from the target value for certain period of time In this phase of implementation we deal with the dataset by replacing these missing values and thus converting it into non-missing value. These samples or chunk of data is considered as an uncertain sample which is close to decision boundary.

The drift in the data is detected using False Positive and False Negative rate. The proposed model predicts average accuracy and the sample selection with fixed size window and adaptive size window.

From the experimental result shown in Table 1, 2 and 3, it is clear that ensemble is better for all types of chosen sample. In conclusion we can say that, an increase in the sample size, parameter selection, and approximation-based approach leads to a better accuracy in terms of prediction

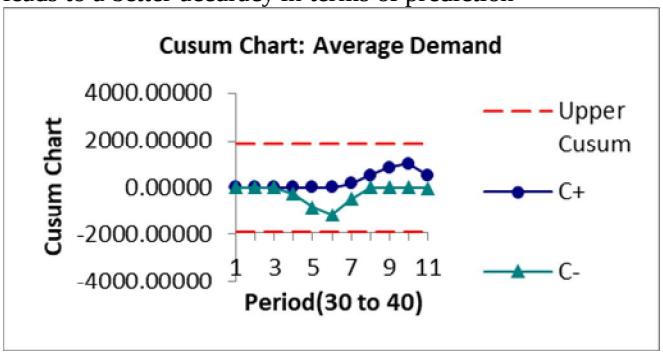

Figure 3: CUSUM chart of Price attribute

Thus, small shifts are identified with CUSM chart and variable size samples are given to single and ensemble classifier.

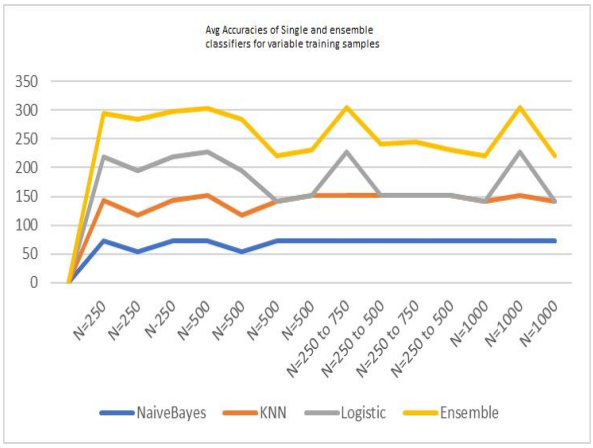

Figure 4: Average Accuracy of Single Vs Ensemble

Figure 4 shows the comparison of classifiers with ensemble classifiers. Variable size samples generates a diversified training data using statistical approaches . In Table 3 the average accuracy of diversified samples .is calculated.

Table 3: Accuracy table single Vs Ensemble

Classifier	*Block Size N=50*	*Block Size N=500*	*Block Size N=1000*
J48	90.9051	90.9051	80.4749
Logistic	89.1248	89.1248	76.0979
Random Forest	95.9214	95.9214	81.9853
Bagging	89.94294	89.94294	83.3485
Naïve Bayes	75.6542	75.6542	72.9652
Proposed Ensemble	***90.9909***	***90.9909***	**89.8426**

5. CONCLUSION

Thus our approach in contrast to the typical fixed sized window approach, proves that a comparatively better accuracy is achieved if impact of sample size, feature selection is continuously monitored using overlapping training data . Based on above observations, in conclusion, the variable size chunk is obtained by CUSUM charts which contributes for detecting a change in the concept and further improving the classifier accuracy. The impact of different samples summarizes in a way that the drift detection algorithms, the ensemble algorithm outperforms other single classifier with a combination of variable size and fixed window size samples. The model is built with adaptive and fixed window size sample and gives better accuracy. However accuracy can be different with different drift detection algorithms and also for different base learner. Thus with the combine approach of weighted ensemble and Statistical data deviation measures are well handed in concept detection and drift detection methods. The proposed approach works well on concept drift data set.

REFERENCES

1. Eugenio Cesario, Carlo Mastroianni, Domenico Talia. **Distributed volunteer computing for solving ensemble learning problems,** *Future Generation Computer System*, Vol. 54, pp. 68–78, 2016. https://doi.org/10.1016/j.future.2015.07.010

2. J Gama, I Zliobait ,A. Bifet, M. Pecheinzkiy and A. Bouchachia, **A survey on concepts drift adaptation** *ACM Computing Surveys* ,Vol.46, No. 4, pp. 231-238,2014

3. N.Street, Y. Kim, **A streaming ensemble algorithm (SEA) for large-scale classification**. *7th ACM SIGKDD Int. Conf. Knowledge. Disc. Data Min. ACM*, pp.377-382,2017.

4. Hu Jie, Li. Tianrui , W.H.Wang , H Fujita. **Hierarchical cluster ensemble model based on**

a. **knowledge granulation**, *Knowledge-Based System*, Vol. 91(C), pp. 179–188 , 2016. https://doi.org/10.1016/j.knosys.2015.10.006

5. H. Issam Laradji, M. Salahadin ,L. Ghouti. **Xml classification using ensemble learning on extracted features**, *Southeast Regional Conference*, ACM *SE 14*, pp. 1:1–1:6 ,2014.

6. Li Peipei, Wu Xindong, Hu Xuegang, **Mining Recurring Concept Drifts with Limited Labeled Streaming Data**. *ACM Trans. Intell. Syst. Technol.* 3, 2, Article 29, pp. 29:1–29:32 ,2012.

7. N Littlestone, M.K Warmuth. **The weighted majority algorithm** *Information and computation'*, Vol. 108, issue 2 , pp. 212–261,1994. https://doi.org/10.1006/inco.1994.1009

8. Liping Jing, Kuang Tian ,Joshua Z. Huang. **Stratified feature sampling method for ensemble clustering of high dimensional data", *Pattern Recognition***, Vol. 48, issue 11, pp. 3688–3702 .2015

9. M. M. Masud, Q. Chen, L. Khan, C.C. Aggarwal, J. Gao, J. Han, A. Srivastava and Oza. **Classification and adaptive novel class detection of feature- evolving data streams**, *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1484-1497,2013.

10. N.C. Oza, S. J. Russell. **Experimental comparisons of online and batch versions of bagging and boosting**. *7th ACM SIGKDD International Conference Knowledge Discovery*, pp. 176-185,2001.

11. Worapat Paireekreng, Worawat Choensawat . **An ensemble learning based model for real estate project classification**. *Procedia Manufacturing*, pp.3852-3859, November,2015 https://doi.org/10.1016/j.promfg.2015.07.892

12. S. Dongre, L.G. Malik, T. Achamma. **Detecting concept drift using HEDDM in data stream**,

International Journal of Intelligent Engineering Informatics, Vol. 7, issue 3, pp. 164-179 ,2019.

13. H. Senaratne, A. Bröring, T Schreck, D. Lehle **Moving on Twitter: using episodic hotspot and drift analysis to detect and characterize spatial trajectories**. *7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks* , pp. 23-30, 2014.

14. T. Singh, M. Kumari, **Role of Text Pre-Processing in Twitter Sentiment Analysis**, *Twelfth International Multi-Conference on Information Processing (IMCIP), Procedia Computer Science 89 Bangalore, India , pp. 549-554,2016.*

15. Muhammad Umar Chaudhry, Jee-Hyong Lee. **Feature selection for high dimensional data using monte carlo tree search**. *IEEE access, 6,* pp.76036 – 76048, Vol. 6, 2018.

16. J. Plasse, ,Adams, M. Niall . **Multiple change point detection in categorical data streams** ,*Statistics and computing* ,Vol. 29, issue 5, pp. 1109–1125,2019. https://doi.org/10.1007/s11222-019-09858-0

17. K. Nishida , K. Yamauchi, **Detecting Concept Drift Using Statistical Testin**. In: Corruble V., Takeda M., Suzuki E. (eds) Discovery Science. DS 2007. *Lecture Notes in Computer Science, Springer, Berlin, Heidelberg*, Vol. 4755, 2007.

18. Danilo Rafael de Lima Cabral, Roberto Souto , Maior de Barros. **Concept drift detection based on Fisher's Exact test**, *Journal of Information Sciences* , Vol. 442, No. 443, pp. 222-234, 2018.

19. Efraimidis, S Pavlos , **Weighted Random Sampling over Data Streams**. [*online] https://arxiv.org/ , [Accessed July 2015]* https://doi.org/10.1007/978-3-319-24024-4_12

20. Maryam and Noor Setiawan ,A Wrapper Feature Selection Based on Ensemble Learning Algorithm for High Dimensional Data. *Advanced Trends in Computer Science and Engineering* vol.8 no.6, pp.1795-1801, November-December 2019. https://doi.org/10.30534/ijatcse/2019/16862019

21. Maganti Syamala and N.J.Nalini **A Deep Analysis on Aspect based Sentiment Text Classification Approaches** *Advanced Trends in Computer Science and Engineering* vol.8 no.5, pp.1795-1801, September - October-2019. https://doi.org/10.30534/ijatcse/2019/01852019

22. *http://moa.cms.waikato.ac.nz/datasets. [Accessed 2017]*

23. Yange Sun, Zhihai Wang, Haiyang Liu,Chao Du, Jidong Yuan,**Online Ensemble using Adaptive windowing for data stream with concept drift**, *International journal of distributed sensor network, Vol 12, issue 5* Volume 2016 .