International Journal on Recent and Innovation Trends in Computing and Communication
Volume: 4 Issue: 3

ISSN: 2321-8169
404 - 407

Review on Concept Drift Detection Techniques

Aabha Anil Thakre
Department of Computer Science & Engineering
G.H.Raisoni College of Engineering and Technology
Nagpur, India.
e-mail- thakre.aabha10@gmail.com

Prof. Snehlata Dongre
Department of Computer Science & Engineering
G.H.Raisoni College of Engineering and Technology
Nagpur, India.
e-mail- snehalata.dongre@raisoni.net

Abstract **-** Detecting the changes and reacting on them is an interesting research topic in current era. Concept drift detection is comes under data stream mining. Process which takeout information from data stream which continuously generated called data stream mining. Normally in data set the data is stationary but problem arises when data is continuously generated that is data stream. So in that case the detection of concept drift is an important task. There are various techniques for drift detection. This paper focuses on some main technique of drift detection.

Keywords*- Data mining, stream mining, drift detection, gradual drift, sudden drift.*

*****

I. INTRODUCTION

Gaining the knowledge about classifiers from data is one of the main jobs in data mining. Most prior and current research in this field is on the stationary environment where a whole dataset is given to the learning algorithm, these set of data are electronically stored and can be retrieved by algorithm several times if essential. Therefore the mark concept which should be learned are stable.

When data stream generated automatically, the possibility of changes in coming data is more. While investigating the coming data when an unexpected file or data occurs that's called drift. This may be spam or any other kind of data that's why the detection of concept drift is an important factor. There are mainly two kinds of concept drift gradual concept drift and sudden concept drift.

When the changes occur in incoming data is slow then there is gradual concept drift. When the sudden change occurs in incoming data then it is sudden concept drift.

Probability of detecting the sudden concept drift is more as compared to the gradual concept drift.

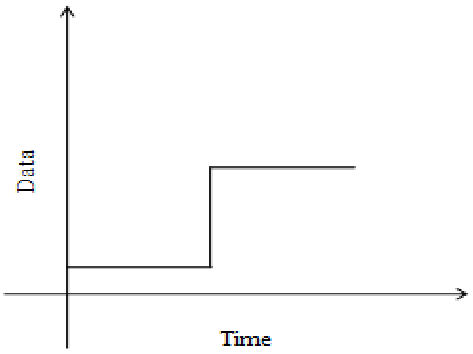

Fig. 1. Sudden Concept Drift

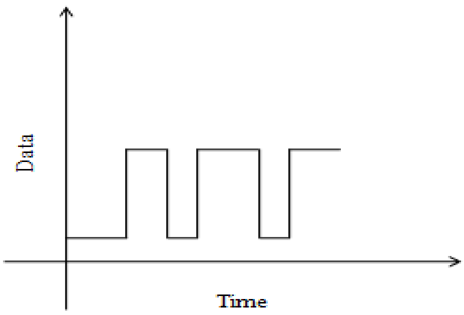

Fig. 2. Gradual Concept Drift

Fig. 1 shows the change which occurs suddenly while Fig. 2 shows the change which occurs slowly.

In some latest appeal learning algorithm works in changeable environments, in which create nonstop production of data. This is called dynamic environment. In this kind of environments, approaching data from a data stream is typed by large volumes of samples or examples, and fast arrival-rate, which requires sudden and actual response. Contrasted to static environments, the refinement of data streams gives new requirements for algorithms, for example bounds on memory handling, limited learning, and checking time, and one scrutinize of incoming samples, due to the changeable nature of data streams, the mark concepts tend to turn or change over time in a function which is called concept drift. Concept drift happens when the concept about samples are being assembled that swings from time to time after a merest stability period.

Let us consider simple example, if we take the online shopping data where data is continuously generated in the form of stream. If the incoming data to our system is of various kinds of foods, that means our machine is train only for detecting the kinds of food. But if in that stream the data related to other field such as menswear comes then the system is not able to recognize that kind of information, so for detecting that kind of things the drift detection techniques available. Our next section covers various drift detection techniques.

II. Related work :

There are various types in which the coming data changes such as Sudden, Gradual. There are distinct algorithm which will helpful for detect concept drift and another advantage is that the algorithm also have the ability of learning with the concept drifting data. The paradigm is study from new characteristic as well as amnesic the previous attribute. Because of this it can skillfully control memory management.

This section consists of mainly two parts one is main Drift Detection Techniques and another is various strategy for detecting the drift.

A. *Drift Detection Techniques*

a. *DDM*

There are various approaches that focus on the errors which produced by the learning model during prediction. The drift detection method (DDM) given by Gama, J., Medas, P., Castillo, G., Rodrigues DDM, as in [1] method uses a binomial distribution. This distribution based on the probabilistic approach. At every point the sequence is sampled. And the number of misclassified elements gives the error rate. It assumes states the PAC learning model given in [2], in PAC it state that, if the number of example exceeds the error rate of the algorithm decreases for stationary environment. A slight change in the error of the algorithm, state that the distribution is changing and, therefore the present decision model may be inappropriate DDM approach is good while detecting sudden changes and, but it has difficulties when the change is slowly gradual. In this situation, the samples will be preserved for more time, the level of drift can take too much time to check and the examples memory can be exceeded.

b. *EDDM*

To overcome the drawback of DDM the new approach comes in existence known as EDDM (Early Drift Detection Method), this method has been created to increase efficiency and improvement of the detection of drift which changes occurred gradually. Simultaneously, it keeps a better performance with sudden concept drift. The main concept behind EDDM is, it assume the range among two error classification rather than concentrated upon the number of errors. The method follows the standard hypothesis in [3].

When there is amount of minimum 30 classification errors, this technique searches for concept drift. It will check 30 error classifications and after occurrence of 30 errors this method uses threshold to detect the concept drift. The Early Drift Detection Method selects 30 classifications of errors because it use distance between two uninterrupted errors and compare that with next distribution in order to check the differences.

c. *Drift Detection based on Diversity*

This technique is mainly based on ensemble diversity. Diversity of an ensemble classifier is an important factor as it will help in determining the error generation. The information given in [4], first ensemble diversity analysis related to online learning when the concept drift is present. For detecting the drift, in [5] & [6], state that different diversity levels are required. In order to get better generalization on the old or new concepts and automatic detection of diversity can aid to decrease the early growth in error which is caused by drift, but it does not provide any fast revival from drifts. The main problem in this technique is, the ability of ensemble with the drift may not be fully used by current ensemble approaches yet, because they do not boost-up different levels of diversity in different situations.

d. *Resampling*

Resampling mainly focus on detection in streaming data and change in prediction loss concept drift. While taking the term resampling the definition of drift slightly diverge from typical definition of concept drift, which identify drifts only with variation in the target concept. Resampling connect a concept drift also to the learning algorithm. Resampling is based on the random permutation of examples.

Their main contribution is a general approach for detection of changing stream in a prediction problem for different kinds of concept drifts. Their detection mechanism relies on the idea of random permutations of the samples, which produce various training-testing splits from the stream of data. This proposed approach given in [7], is more robust for noisy changes.

e. *HDDM*

HDDM based on the idea of statistical process control. It provides three states. Stable, warning and drift level. In HDDM [8], they define two different values one for warning level and another for drift level. Stable state will show that there is no change in the input data stream which is generated, when it senses that may be a change may appear then that will shown by warning level and when the drift is occur shown by drift level. This technique assumes that if the warning level is exceed then maybe there is possibility of drift. And according to that the alternative classifier is updated, so when it confirms as the drift is occurred the alternative classifier will take the place of old one classifier.

f. *DCDD*

Another approach for detecting the concept drift is DCDD [9]. It follows a simple sequence. It uses short term memory for storing the training examples. When the online classifier rebuild, the short term memory enhance the predictive accuracy immediately after that. This method focuses on the length of window size, base learner algorithm. Two levels are defined in this method warning level and drift level. This technique mainly use for email spam detection.

g. *EDIST*

Error distance based approach for drift detection i.e. EDIST it uses the idea of standard Early Drift Detection Method, the Early Drift Detection Method check the error between two uninterrupted errors of classification, and the difference between early drift detection and EDIST is that EDIST takes two data window. EDIST traces concept drift by maintaining two windows. One is global sliding window and another window uses to store the present example. EDIST uses the standard hypothesis given in [10]. While using this hypothesis, noticed that r and s parameters affects on the toughness of EDIST towards false alarm and noise.

h. *Hierarchical CDT*

The reduction of false positive rate in concept drift called hierarchical CDT. H-CDT has two levels. First level operate CDT test which gives an alarm of drift. It may be either real concept drift or false positive alarm & gives the value stated in [11]. In second level, Based on the value of t' this level CDT divide the data stream in to two parts, before & after change. For calculating this it will make a use of hotelling test. The Hotelling test break or take the concept drift detection which is proposed by CDT at first level. And its result is the final to be considered. In other words Hotelling test check, if the present contents in stream before and after the change are the same or not.

B. *Various Strategies for detecting drift*

i. *Windowing techniques*

Windowing technique uses the concept of sliding window. This is well liked tactic to hand out with time emerging data. Sliding Window give permission to limiting the samples which is going to learn as well as it flush out the old data and add new data, so the paradigm gain the newest idea. The basic idea behind this windowing algorithm is very clear. In windowing technique each sample update the window rather than updating by classifier. A double window strategy shows in [12].

j. *Weighted Windows*

Weighted window uses decay function. This decay function will helpful for flushing out previous samples which are of no need. Decay function allocates some weight to samples in window. In weighted window previous samples gains little weights and it has less privilege than new sample, that's why they are less important than classifier. Cohen and Strauss provide use of decay functions for data stream aggregates calculation, given in [13].

k. *FISH*

Zliobaite proposed a technique called FISH algorithm. FISH make a use of time and space uses time and space likenesses between samples for creating window dynamically. The standard hypothesis is given [14]. In which the selection of samples is based on distance measure. To properly manage the balance in between time and distance, the distance and time should be normalized. The next part of FISH is FISH2, FISH3. Time/space proportion is takes as parameter in FISH2 algorithm. While in FISH3 for each incoming new instance FISH3 algorithm computes time/space proportion and size of window differently.

So the Family of FISH algorithm is very useful while classification when the dynamic windowing technique is necessary in drift detection.

l. *ADWIN*

When there is occurrence of sudden drift ADWIN is appropriate. It is based on sliding window approach. ADWIN made up of sliding window having size w which stored freshly occurring samples. ADWIN consider two sub-windows having different averages so that predicted output values are different. It makes use of pseudo-code for checking average in sub-windows. All the calculation is based on the threshold value and it'll calculate by using Hoeffding bound, given in [15] & [16].

m. *SEA*

Street and Kim proposed an ensemble method which changes its arrangement to acknowledge the change that method is Streaming Ensemble Algorithm (SEA). SEA use substitution plan for runt classifier. For that SEA uses two approaches one is accuracy and another is diversity. SEA divides incoming data stream in to chunks. And the new classifier is trained in each chunk. After the completion of process the weaker classifier is flushed out and new one is preferred. Accuracy Weighted Ensemble and accuracy diversified Ensemble are some of ensemble algorithm, given in [17], and Accuracy updated Ensemble approach is given in [18].

n. *ACE*

Nishida proposed a hybrid method in which drift detector will help the data chunk ensemble called Adaptive Classifier Ensemble (ACE), ACE react on sudden drift by tracking error rate of classifier per each incoming sample as well as it gradually rebuild the classifier with big chunks of samples as in [19]. ACE is based on the statistical decision theory which provides rank based selection.

o. *ECDD*

G. J. Ross, N. M. Adams, D. Tasoulis, D. Hand projected a drift detection method which is focus on Exponentially Weighted Moving Average (EWMA), which applied for differentiate raise in arbitrary variables, given in [20]. In EWMA, the incorrectly ordering before the changing point and standard deviation of data stream are known. Where as in ECDD, approximation of improvement and dissatisfaction denoted as(1 and 0).

p. *STEPD*

Nishida., stated a method which is based on statistical test called Statistical Test of Equal Proportions (STEPD). In STEPD, the accuracy of recent classifier will be as equal to overall accuracy is expected where as if the environment is stationary and concept is changing which is given in [21]. A test called chi-square is fulfilled by rectification of measurement and its class. On the other hand, if it is not on the centrality level, then the unacceptable opinion is prohibited. The threshold value of warning and drift level are furthermore calculated like shown in DDM EDDM etc.

q. *DOF*

The method which recognizes drift by depuration of data chunk by chunk is DOF. It is proposed by Sobhani P. and Beigy H. In DOF, for each occurrence it computed the closest neighbor in every previous batch and current batch and comparing their corresponding labels. According to the index occurrence from previous or current batch the distance map is generated by summing adjacent neighbor. And based on distance map the degree of drift is also calculated in this method. According to standard deviation and the average of degree of drift, if the current output is far from average which is greater than standard deviation a concept drift is generated. This DOF algorithm is efficient in case of separated and balanced classes given in [22].

III. CONCLUSION:

For solving drift detection problem it required learning algorithm and hoeffding bounds algorithm. When change occurs in data samples drift detection method works. During the classification process the accuracy of the algorithm is improved. Since drift detection and classification algorithm adapted while change occurs in data stream.

This Paper describe about concept drift and its types, it's requirement and details of changes. There are various techniques DDM, EDDM, DDD, Drift detection through resampling, HDDM, DCDD along with that some windowing techniques, ensemble methods, and statistical method are made known in this paper. DDM is very suitable for sudden concept drift but it lack in detecting the slow change in data stream so the next technique EDDM will help to overcome this drawback. In determining the error generation diversity plays an important role.

REFERENCES:

[1] Gama, J., Medas, P., Castillo, G., Rodrigues, "P.: Learning with drift detection". Lecture Notes in Computer Science 3171 (2004)

[2] Mitchell, "T.: Machine Learning". McGraw Hill (1997)

[3] M. Baena-García, J. D. Campo-Avila, R. Fidalgo, A. Bifet, R. Gavaldà, and R. Morales-Bueno. "Early drift detection method", In Fourth International Workshop on Knowledge Discovery from Data Streams, 2006.

[4] Leandro L. Minku, and Xin Yao," DDD: A New Ensemble Approach for Dealing with Concept Drift," IEEE Transaction On Knowledge Data Engineering., Vol. 24, No. 4, April 2012, Pp 619-633.

[5] F.L. Minku, A. White, and X. Yao, "The Impact of Diversity on Online Ensemble Learning in the Presence of Concept Drift," IEEE Transaction Knowledge and Data Eng., vol. 22, no. 5, pp. 730-742, http://dx.doi.org/10.1109/TKDE.2009.156, May 2010.

[6] L.I. Kuncheva and C.J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," Machine Learning, vol. 51, pp. 181-207, 2003.

[7] Maayan Harel, Koby Crammer, Ran El-Yaniv, Shie Mannor, "Concept Drift Detection Through Resampling,".

[8] Isvani Frıas-Blanco, Jose del Campo-Avila, Gonzalo Ramos-Jimenez, Rafael Morales-Bueno, Agustın Ortiz-Dıaz, and Yaile Caballero-Mota, "Online and Non-Parametric Drift Detection Methods Based on Hoeffding's Bounds," IEEE Transaction On Knowledge Data Engineering., Vol. 27, No. 3, March 2015, Pp 810-823.

[9] L. Nosrati, A. Nemaney Pour, "Dynamic Concept Drift Detection for spam Email Filtering," in ACEEE Int. J. on Network Security, Vol. 02, No. 04, October 2011.

[10] Imen Khamassi, Moamar Sayed-Mouchaweh,"Drift detection and monitoring in non-stationary environments". IEEE 2014.

[11] Cesare Alippi, Derong Liu, Dongbin Zhao, and Li Bu, "Detecting and Reacting to Changes in Sensing Units: The Active Classifier Case," IEEE Transactions On Systems, Man, And Cybernetics: Systems, Vol. 44, No. 3, March 2014.

[12] Qun Zhu, Xuegang Hu, Yuhong Zhang, Peipei Li, Xindong Wu, "A Double-Window-based Classification Algorithm for Concept Drifting Data Streams", IEEE International Conference on Granular Computing, 2010.

[13] Edith Cohen and Martin J. Strauss. "Maintaining time decaying stream aggregates". J. Algorithms, 59(1):19–36, 2006.

[14] Indr˙e ˇZliobait˙e. "Adaptive training set formation". PhD thesis, Vilnius University, 2010.

[15] Albert Bifet and Ricard Gavald`a. Kalman filters " adaptive windows for learning in data streams". In Ljupco Todorovski, Nada Lavrac, and Klaus P. Jantke, editors, Discovery Science, volume 4265 of Lecture Notes in Computer Science, pages 29–40. Springer, 2006.

[16] Albert Bifet and Ricard Gavald`a. "Learning from timechanging data with adaptive windowing." In SDM. SIAM, 2007.

[17] W. Nick Street and YongSeog Kim. "A streaming ensemble algorithm (sea) for large-scale classification", In KDD, pages 377–382, 2001.

[18] Dariusz Brzezinski and Jerzy Stefanowski, "Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm", IEEE Transaction On Neural Networks And Learning Systems, Vol. 25, No. 1, January 2014.

[19] K. Nishida, K. Yamauchi, and T. Omori, "ACE: Adaptive classifiers- ensemble system for concept-drifting environments", in Proc. 6th Int. Workshop Multiple Classifier System., 2005.

[20] G. J. Ross, N. M. Adams, D. Tasoulis, D. Hand(2012), "Exponentially weighted moving average charts for detecting concept drift", International Journal Pattern Recognition Letters, 191-198.

[21] K. Nishida(2008), "Learning and Detecting Concept Drift", A Dissertation: Doctor of Philosophy in Information Science and Technology, Graduate School of Information Science and Technology, Hokkaido University.

[22] Sobhani P. and Beigy H.(2011), "New drift detection method for data streams", Adaptive and intelligent systems, Lecture notes in computer science, Vol. 6943, pp. 88–97.