[1]Meenakshi A. THALOR

ANALYSIS OF DIFFERENT DRIFT DETECTION TECHNIQUES ON DATA STREAM

[1]Department of Information Technology, AISSMS Institute of Information Technology, Pune, Maharashtra, INDIA

Abstract: Traditionally, data mining assumes that training and testing dataset are produced from a single, stationary and hidden function. Its means that, the data used at testing time is generated from same function as from training time. In the data stream the above-mentioned assumption is not true that is the source which generates examples at training time need not be the same source which generates examples at testing time. This paper highlights the change of source of data which can be abrupt, gradual, incremental or reoccurring. At the end this paper provides analysis of drift detection methods on abrupt, gradual or incremental drifted data stream.

Keywords: Data Stream, Concept Drift, Drift detection

DATA STREAM CLASSIFICATION

Data stream is continuously arriving sequence of data example available at specific rate, unlimited in size because of that some of data examples are discarded once analyses are done; data examples possess continuously evolving pattern and bustiness as generated from different sources. Traditional data mining approaches is not suitable to data stream because of aforementioned characteristics of data stream. Incremental models [1] are used to handle data stream where the data is fetched at time *t* rather than retrieving all the data at the beginning of training. Model process each instance with in a constant time and takes only one look on to the data rather than multiple passes. Incremental model makes use of fixed amount of memory for storage and always ready to predict at any time.

Classification helps in decision making by providing the class labels for given data using historical records. Figure 1 shows the classification process on data streams [1] where data chunks C_1; C2; C3…….C_i arrive one after one. A data chunk consists of bunch of instances. In two class problem, each chunk contains some class1 instances and some class2 instances.

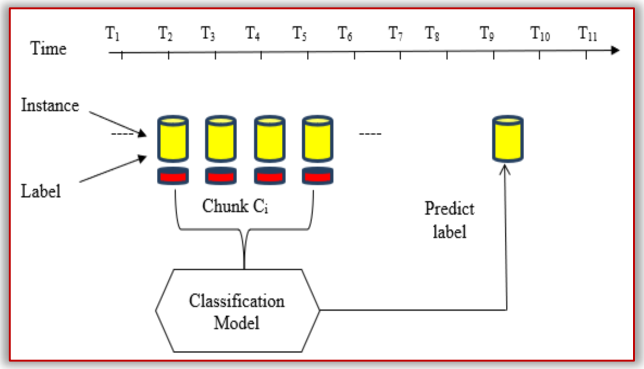

Figure 1: Classification of Data Stream

Suppose C_1; C2; C3…….C_i are labeled. At the time stamp T_9, when an unlabeled chunk C_{m+1} will arrive, the classification model will provide the class labels of instances available in C_{m+1} on the basis of training data. If the prediction given by classification model for chunk C_{m+1} is correct, the C_{m+1} chunk can join the training set, resulting in more and more availability of training data.

A storage constraint makes it important to carefully select instances which can represent the current distribution. Most studies on data stream mining assume relatively balanced and stable distribution of data in streams. However, most of real time applications involve concept-drifting data streams with unbalanced distributions [2], because of these issues the classification of data stream is a prominent area of research.

Concept Drift

In the data stream, the source which generates examples at training time need not be the same source which generates examples at testing time and this change of source of data is called as concept drift. Thus, data from the previous source may be useless or irrelevant for the current context.

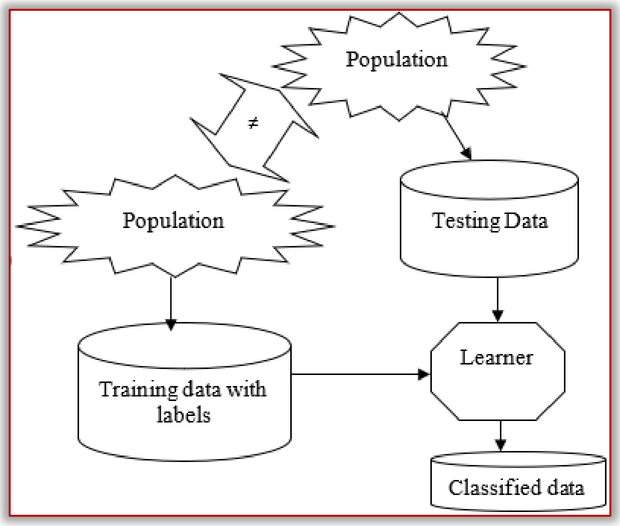

Figure 2: Learning on Non stationary Data Stream

Figure 2 depict the supervise learning on non-stationary data stream where the population data generated at training time is not equivalent to population data at testing time and represented by following equation (1).

$$P_{tr}(Y|X) \neq P_{tst}(Y|X) \text{ and } P_{tr}(X) = P_{tst}(X) \quad (1)$$

The change in source can be categorized in following ways:

SUDDEN CONCEPT CHANGE

In nonstationary data stream, data may be generated from different source, sudden concept change occurs at a point in time when data source changes from one concept to another concept. In figure 3 from timestamp T_2 to T_4 some instances are coming from source S1 and represented as concept 1 and from timestamp T5 to T8 some instances are coming from source S2 and represented as concept 2. At timestamp T_5 the concepts suddenly change from concepts 1 to concept 2 which is called as concept drift as concept 1 instances are generated by source1 and concept 2 are generated by source 2 and hence their data distribution is different as shown by Mean axis. In order to represent different concepts different colors are used and change of color shows drift occurred.

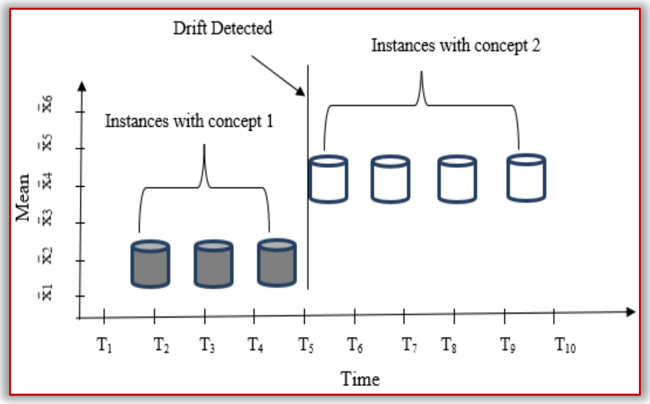

Figure 3: Sudden Concept Change

INCREMENTAL CONCEPT CHANGE

An incremental concept occurs when there are multiple concepts in data stream which are generated from multiple sources and the difference among the multiple sources is very small. In figure 4 concepts are changing as follows:

- Time stamp T_4: concepts change from concepts 1 to concept 2
- Time stamp T_5: concepts change from concepts 2 to concept 3
- Time stamp T_7: concepts change from concepts 3 to concept 4

All these concepts are generated from different sources and hence their data distribution is different as shown by Mean axis.

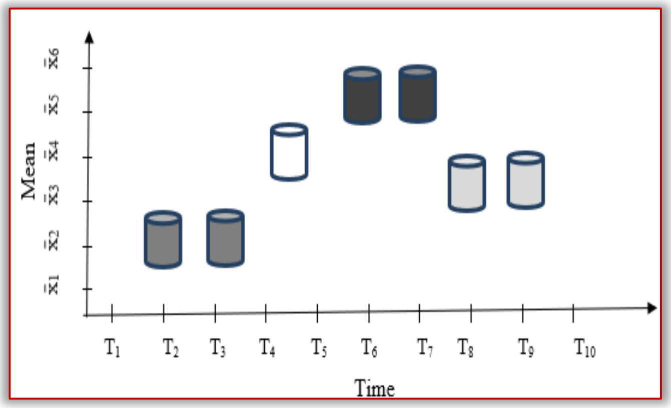

Figure 4: Incremental Concept Change

GRADUAL CONCEPT CHANGE

When two or more data sources generate data after some time stamp, it is called gradual drift. In figure 5 concepts are changing as follows:

- Time stamp T_4: concepts change from concepts 1 to concept 2
- Time stamp T_5: concepts change from concepts 2 to concept 1
- Time stamp T_7: concepts change from concepts 1 to concept 2

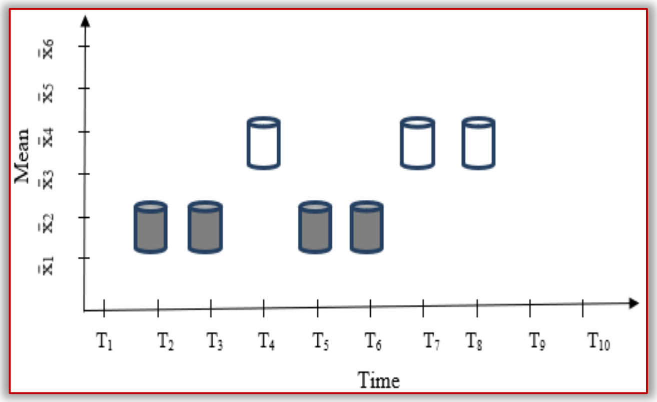

Figure 5: Gradual Concept Change

REOCCURRING CONCEPT CHANGE

Reoccurring concepts occurs when same data is generated over a period of time using different data sources (similar to incremental and gradual drift). It is different from incremental and gradual drift as same sources are used to generate data in near future. In figure 6 concepts are changing as follows:

- Time stamp T_3: concepts change from concepts 1 to concept 2
- Time stamp T_5: concepts change from concepts 2 to concept 1
- Time stamp T_7: concepts change from concepts 1 to concept 2
- Time stamp T_9: concepts change from concepts 2 to concept 1

This shows a pattern or repeated behavior as concept pattern is reoccurring after some time.

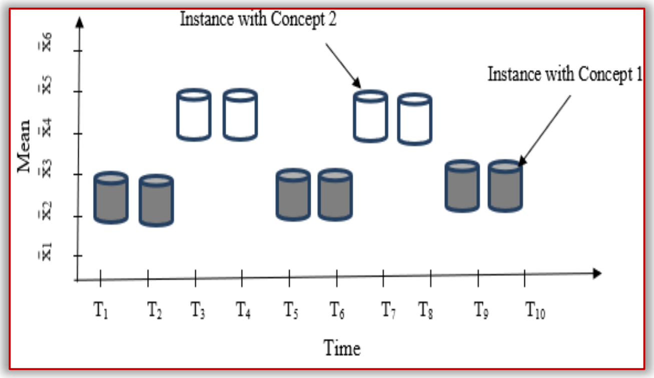

Figure 6: Reoccurring Concept Change

DRIFT DETECTION METHODS

Drift Detection Method (DDM) was introduced by Gama et al. [3] which based on Binomial Distribution. In addition to classification error rate, authort defines two levels i.e. the warning level and the alarm level. In time series data, if at i_w the error reaches at the warning level and at i_d it represents the alarm level then it is considered the drift has occurred and drift detector starts the training process with data from i_w. DDM gives better results on data streams which possess sudden drift change because gradually drifts can escape from detector without activating the alarm level.

DDM is further modified as EDDM which was proposed by Baena-Garcia et al. [4] works on distance error rate rather than classification error rate and make use of warning and alarm level thresholds. Once an alarm level is reached, there is need to construct a new model on data from warning level. It make use of a threshold value 30 to search for a concept drift.

ADWIN is ADaptive WINdowing algorithm proposed by Bifet et al [5] which make use of variable size sliding windows. It uses averaging of elements in the window to detect the drift. The size of window can grow or shrink as and when no drift or change is detected.

HDDM is Hoeffding Drift Detection Methods proposed by by Frias-Blanco et al. [6] where author conducted two tests named as HDDMA-test and HDDMW-test. HDDMA-test compares the moving averages to find the drift and EMWA forgetting scheme [7] is used to weight the moving averages. HDDMA-test and HDDMW-test make use of Hoeffding's inequality [8] to set an upper limit to the level of difference between averages. The weighted moving averages are compared to detect the drift in stream. Results show that HDDM is better to detect abrupt and gradual drifts.

CUSumDM [9] represents Cumulative Sum Detection method which was introduced in Biometrika in year 1954, E.S. Page. It is used for change detection using sequential analysis technique.

Recent drift detection method [10] make use of dynamic dynamic classifier selection in order to detect drift while [12] consider variety of measure for drift detection in data stream.

EXPERIMENTS AND RESULTS

This paper provides overview of concept drift in data stream by using various drift detection mechanisms. Drift is introduced in synthetic dataset using gradual and sudden generators available in MOA framework. Using synthetic datasets one can analyse how different methods deal with the different types of drift as beginning and end position of drift known in advance.

Total 400000 instances are created and drift is introduced in data stream using SEA generator and hyperplane generator. Experimental results of table 1, table 2 and table 3 are carried out on the abrupt, gradual and incremental drifted data stream respectively using 6 drift detection mechanisms as mention in section 2. Naïve Bayes classifier is used for the learning purpose.

Table1: Results of Drift detection mechanism on Sudden Drift

Drift detection Method	Accuracy (%)	#drifts detected
DDM	80.67	1
EDDM	80.67	2
CUsumDM	81.06	5
HDDM_A	82.65	2
HDDM_W	84.52	20
ADwin	84.84	10

Table 2: Results of Drift detection mechanism on Gradual Drift

Drift detection Method	Accuracy (%)	#drifts detected
DDM	80.69	1
EDDM	80.68	2
CUsumDM	81.07	3
HDDM_A	81.68	3
HDDM_W	83.72	19
ADwin	83.63	49

Table 3: Results of Drift detection mechanism on Incremental drift

Drift detection Method	Accuracy (%)	#drifts detected
DDM	79.36	1
EDDM	79.36	2
CUsumDM	79.36	1
HDDM_A	79.51	1
HDDM_W	82.11	34
ADwin	82.45	13

Experiment results conclude with following observations that HDDM_W and ADwin detects more drift in data stream with high accuracy whereas DDM, EDDM, CUsumDM and HDDM_A provide approximate similar results. The accuracy can be further enhanced by using ensemble [10][11].

CONCLUSION

The problem of concept drift with its need and type is elaborated in this paper. The several concept drift detection methods like DDM, EDDM, ADWIN, HDDM and CUSumDM and ECDD are discussed with their experimental results.

Results shows that Naïve based classifier provide better accuracy in case of concept change. The accuracy can be further enhanced by using multiple classifiers and results can be shown on different applications.

References

[1] Thalor, M.A.; Patil, S.T.: Incremental learning on non-stationary data stream using ensemble approach, International Journal of Electrical and Computer Engineering, 6(4), 1811–1817 (2016).

[2] Thalor, M.A.: Patil, S.T.: Learning framework for non-stationary and imbalanced data stream, International Journal of Engineering and Technology, 8(5), 1942–1945 (2016).

[3] Gama, J.; Medas, P.; Castillo, G. and Rodrigues, P.: Learning with Drift Detection, Lecture Notes in Computer Science, 3171, 286-295 (2004).

[4] Baena-Garcia,M.; Campo-Avila ,J.; Fidalgo, R.;Bifet ,A.;, Gavaldµa R. and Morales-Bueno R: Early Drift Detection Method, IWKDDS, 77–86 (2006).

[5] Bifet,A. :Adaptive Learning and Mining for Data Streams and Frequent Patterns, Doctoral Thesis.(2009).

[6] Frias-Blanco, I.; del Campo-Avila, J.; Ramos-Jimenez, G.; Morales-Bueno, R.; Ortiz-Diaz, A.; Caballero-Mota, Y.: Online and Non-parametric Drift Detection Methods based on Hoeffding's Bounds. IEEE Transactions on Knowledge and Data Engineering 27(3), pp. 810–823 (2015)

[7] Ross, G. J., Adams, N. M., Tasoulis, D. K., Hand, D. J.: Exponentially Weighted Moving Average Charts for Detecting Concept Drift. Pattern Recognition Letters, 33(2), pp. 191–198 (2012)

[8] Hoeffding, W.: Probability Inequalities for Sums of Bounded Random Variables. American Statistical Association 58 (301), pp. 13–30 (1963)

[9] Granjon, P.: The CuSum algorithm-a small review (2013).

[10] Pinage, F.A., Santos, E.M., & Gama, J.: A drift detection method based on dynamic classifier selection. Data Mining and Knowledge Discovery, 34, 50-74 (2019).

[11] Mahdi, O.A., Pardede, E., Ali, N., & Cao, J.: Diversity measure as a new drift detection method in data streaming. Knowl. Based Syst., 191, 105227(2020).

[12] Thalor, M.A.; Patil S.T.: Review of ensemble-based classification algorithms for nonstationary and imbalanced data. IOSR Journal of Computer Engineering, 16(1), 103–107(2014).

[13] Thalor, M.A.; Patil, S.T.: Learning on high frequency stock market data using misclassified instances in ensemble. International Journal of Advanced Computer Science and Applications, 7(5):283–288 (2016).

ACTA TECHNICA CORVINIENSIS – Bulletin of Engineering
ISSN: 2067-3809
copyright © University POLITEHNICA Timisoara,
Faculty of Engineering Hunedoara,
5, Revolutiei, 331128, Hunedoara, ROMANIA
http://acta.fih.upt.ro