International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 09 | Dec-2015 www.irjet.net p-ISSN: 2395-0072

Survey on Method of Drift Detection and Classification for time varying data set

K. Wadewale[1], S. Desai[2]

[1] *PG scholar, Department of Computer Engineering,*
MIT College of Engineering,
Pune, Maharashtra, India

[2]*Associate Professor, Department of Computer Engineering,*
MIT College of Engineering, Pune,
Maharashtra, India

Abstract - ***The major problem of online learning or incremental learning is that, target function is frequently changing over time. This problem is commonly known as concept drift. Concept drift can be is further complicated if the dataset is class-imbalanced. There are different learning methods presented so far to handle concept drift like rule-based systems, decision trees, Naive Bayes, support vector machines, instance based learning, ensemble of classifiers, etc. This learning method requires further to combined with methods of drift detection in order to constantly monitor the performance of concept drift, however online changes detection was failed. In literature there are many methods presented for learning from data streams and drift detection, but most of methods failed to achieve speed and accuracy due to data inconsistency. In this project our goal is to present efficient method for online and non-parametric drift detection. This proposed method is based on recently presented Hoeffdings Bounds and HDDM. It handles concept drift regardless of the learning model to monitor the performance metrics measured during the learning process, to trigger drift signals when a significant variation has been detected. The existing system however as Naive Bayes classifier are having limitations, there is no scope to improve accuracy of HDDM. The Propose system will be efficiently provide drift detection method for data stream mining to improve accuracy.***

Key Words: *Concept Drift, Hoeffdings Bounds and HDDM, online learning method, classification, data streams.*

1. INTRODUCTION

Today, is a world of innovative technologies, each field is automated. Due to advances in technology, plenty of data is created every second. Examples of such applications include network monitoring, web mining, sensor networks, telecommunications data management, and financial applications [1]. The information needs to be collected and processed, to extract unknown, beneficial and interesting knowledge. But it is impossible to manually extract that knowledge due to the volume and speed of the data gathered. Concept drift happens when the concept about which information is being collected shifts from time to time after a least stability period. This problem of concept drift needs to be considered to mine data with acceptable accuracy level. There are cases of concept drift contain spam detection, financial fraud detection, weather change forecast, customer preferences for online shopping.

2. OVERVIEW

2.1 Problem of Concept Drift

There has been expanded significance of concept drift in machine learning and additionally information mining tasks. Today, data is organized in the form of data streams reasonably than static databases. Also the concepts and data circulations must to change over a long period of time.

2.2 Need for Concept drift adaptation?

The dynamically changing or non-stationary environments, the data distribution can change over time results in to the phenomenon of concept drift [2]. The concept drifts can be rapidly adapted by storing concept descriptions, so that it can be re-examined and reused after words. Therefore, adaptive learning is required to deal with the data in non-stationary environments. When concept drift is detected, the present model needs to be updated to retain accuracy.

2.3 Types of Concept drift

Depending on the relation between input data and target variable, concept change take different forms. Variations

of target concepts are characterized into sudden, incremental, gradual, recurring, blip or noise drifts. Figure 1 contains the six basic types of drifts. The first plot shows sudden changes that instantly and irreversibly change the data cases of respective class. The next two plots (Incremental and Gradual) explain changes that take place slowly over time. Incremental drift occurs when data example gradually changes their values over time, and gradual drift occurs when the change in data example includes the class distribution of various data.

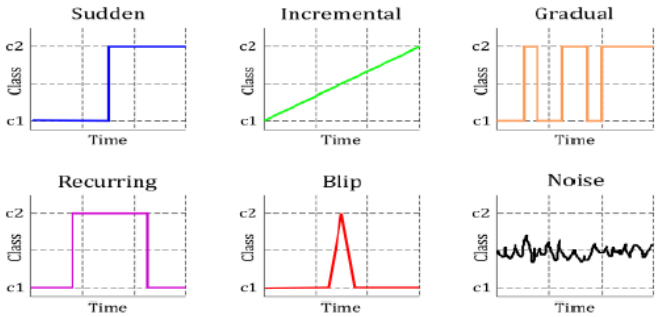

Fig- 1: Types of concept drift

The left-bottom plot (Recurring) signifies changes in data example that are only temporary and are returned after some time. The fifth plot signifies a rare event which can be considered as an outlier in a static distribution. The last plot in Figure 1.signifies random changes in data examples, which would be filtered out efficiently. A good classifier should learn incrementally and adapt to such changes.

2.4 Detecting Concept changes

The ways to monitor concept drift are as given below:

- Concept drift is observed by checking with the data's probability distribution, since it will changes with time.
- One can evaluate whether concept drift has happened, by monitoring and tracking the applicable between various sample characteristics or attributions.
- Concept drifts signifies to changes in features of classification models.
- Classification accuracy to be taken into account while detecting concept drift on a provided data stream. Recall, precision are some of the correctness indicators of classification the timestamp of single sample or block sample to be taken as an additional input attribute, to identify occurrence concept drift. It gives a check on whether the classification rule has become outdated.

3. CONCEPT DRIFT DETECTORS

The section discusses algorithms permitting detecting concept drift, known as the concept drift detectors. it alarm the base learner, that the model should be reconstructed or updated.

3.1 DDM

DDM stands for Drift Detection Method where each iteration of the online classifier calculates the decision class which is either true or false [3]. So for the set of examples error is identified from Bernoulli trials. For each example in data stream we have to update two register to keep track of error rate first is *pmin* and Secondly the *smin*. This two are used to identify warning level condition and alarm level condition. Whenever there is warning level reach examples are remember in the separate window, and if alarm level reaches the previously learned classifier is dropped and new is accepted from the example stored in separate warning window.

3.2 EDDM

EDDM is modification of DDM proposed by Baena-Garcia et al. In this algorithm use the same warning-alarm mechanism which is recommended in DDM, but it uses the distance error rate instead of classifier's error rate. EDDM performs better in the case of gradual drift but it is more sensitive to noise [4].

3.3 ADWIN

Bifet et al. proposed this method which uses sliding windows of variable size, that are recomputed online according to the rate of change detected from the data in these windows [5]. The window (W) is dynamically enlarged when there is no clear modification in the context, and shrinks it when a modification is detected. Accordingly, ADWIN gives rigorous assurances of its performance, in the form of limits on the rates of false positives and false negatives. ADWIN works only for one-dimensional data. A separate window should be maintained for each dimension, for n-dimensional raw data, which results in handling more than one window.

3.4 Exponentially weighted moving average for Concept Drift Detection (ECDD)

Ross et al., proposed a drift detection technique taking into account Exponentially Weighted Moving Average (EWMA), utilized for distinguishing an increment as a part of the mean of an arrangement of arbitrary variables [6]. In EWMA, the likelihood of erroneously ordering an occasion before the change point and the standard deviation of the

stream are known. In ECDD, the estimations of progress and disappointment likelihood (1 and 0) are figured online, taking into account the arrangement precision of the base learner in the real example, together with an estimator of the normal time between false positive detections.

3.5 Statistical Test of Equal Proportions (STEPD)

The STEPD proposed by Nishida et al., expect that 'the accuracy of a classifier for recent W example will be equivalent to overall accuracy from the earliest of the learning if the target concept is stationary; and a huge decline of recent accuracy recommends that the concept is changing'[7]. A chi-square test is performed by processing a measurement and its quality is contrasted with the percentile of the standard ordinary conveyance to get the watched importance level. On the off chance that this worth is not exactly a centrality level, then the invalid speculation is rejected, expecting that an idea float has happened. The warning and drift thresholds are likewise utilized, like the ones exhibited by DDM, EDDM, PHT, and ECDD.

3.6 DOF

The methodology proposed by Sobhani et al. identify drifts by processing data chunk by chunk , the nearest neighbor in the earlier batch is computed for each instance in the present batch and comparing their equivalent labels. A distance map is generated, associating the index of the instance in the earlier batch and the label computed by its adjacent neighbor; degree of the drift is computed based on the distance map. The average and standard deviation of all the degrees of drift are computed and, if the present value is away from the average greater than standard deviations, a concept drift is raised, where s is a factor of the algorithm. This algorithm is more effective for the problems with well separated and balanced classes [8].

3.7 HDDM

We have presented a simple method to differentiate between three separate states: STABLE, when it seems to be no change; WARNING, when it seems that a probable concept drift may appear; and DRIFT, when the drift is clearly recognized [9]. The information provided by the method in the variable STATE can be utilized in many ways and our does not limit the actions to be performed when warning or drift states are identified. However, one of the most direct usages is the following: a) if the warning level is exceeded a possible drift will reach and, consequently, new detected examples can be buffered and utilized to train an alternative classifier; b) when the drift signal is triggered a hypothetical alternative classifier could replace the old one to adapt learner using the buffered examples. Statistical test (A-test or W-test) is to estimate the actual state (STABLE, WARNING or DRIFT) of the change detector from αw and αD over the samples X1, X2 . . . Xcut and Xcut+1. . . Xn. This way, if the null hypothesis is rejected with the size αW the current status is set to WARNING. Similarly, if the null hypothesis is rejected with size αD, the change detector reaches DRIFT level and all the counters are reset. Otherwise the null hypothesis is granted and the present status is set to STABLE. We have called this online change detector HDDM because it is similar to DDM but uses in its place the Hoeffing's inequality for the two-sample statistical test.

4. CONCEPT DRIFT HANDLING

Recommends to group ensemble strategies for changing environments as follows [10]:

- Dynamic combiners (Horse racing): component classifiers are trained and their combination is changed using forgetting process.
- Updated training data: component classifiers in the ensemble are generated incrementally by incoming examples.
- Updating the ensemble member: ensemble members are restructured online or retrained with blocks of data.
- Structural changes of the ensemble: ensemble members are re-evaluated and the worst classifiers are updated or replaced with the classifier trained on the most recent examples, with any concept change.
- Adding a new feature - The attributes used are changed, as an attribute becomes significant, without redesigning the collaborative structure.

The approaches to handle concept drifts contain single classifier and ensemble classifier approaches. The single classifiers are traditional learners that were modelled for immovable data mining and have the qualities of an online learner and forgetting mechanism. Basically, ensemble classifiers are sets of single classifiers whose individual decisions are gathered by a voting rule. The ensemble classifiers provide improved classification accuracy as compared to the single classifiers combined decision. They have a normal way of adapting to concept changes due to their modularity [9].

4.1 Streaming Ensemble Algorithm (SEA)

The SEA, proposed by Street and Kim, changes its in structure based on concept change [11]. It's a heuristic replacement approach of the weakest base classifier based on correctness and diversity. The collective decision was based on majority voting and base classifiers unpruned. This algorithm works best for at most 25 components of the ensemble.

4.2 Accuracy Weighted Ensemble (AWE)

In SEA, it's critical to properly define the data chunk size as it determines the ensembles flexibility. The algorithm AWE, recommended by Wang et al., trains a new classifier C' on each incoming the data chunk and use that chunk to calculate all the existing ensemble members to select best component classifiers. AWE is the best suited for large data streams and works well for the periodic and other drifts.

4.3 Adaptive Classifier Ensemble (ACE)

To overcome AWE‟s gradual drift reactions, Nishida proposed a hybrid approach where a data chunk ensemble is aided by the drift detector, called Adaptive Classifier Ensemble (ACE), aims at reacting to abrupt drifts by tracking the classifier's error rate per each incoming example, while gradually reconstructing a classifier ensemble with large chunks of examples.

5. PROPOSE PLAN FOR WORK

From the above discussion, it has been observed that there is a requirement of quick and efficient classifier to define concept drift and classify data examples precisely. This section provides a design flow and description of primary model for classification and drift finding. The proposed model aims to eliminate the problems related to inefficiency in accurately classifying the data examples in presence on concept drift as the base classifier was not able to learn the old concept well and thus there is need of completely new classifier is required to train on new concept in order to generate negative instances positive and detect drifted data along with classification based on completely new concept. The model aims not only to precisely classify the data but also detect drifted data correctly. Figure show the work flow of the proposed system.

The steps that are followed to obtain accurately classified data along with drifted data are:

1) The system evaluates dataset properties.
2) The redundant and inappropriate attributes are removed.
3) Minority class is identified and iteratively the boundaries are refined to obtain precisely classified instances.

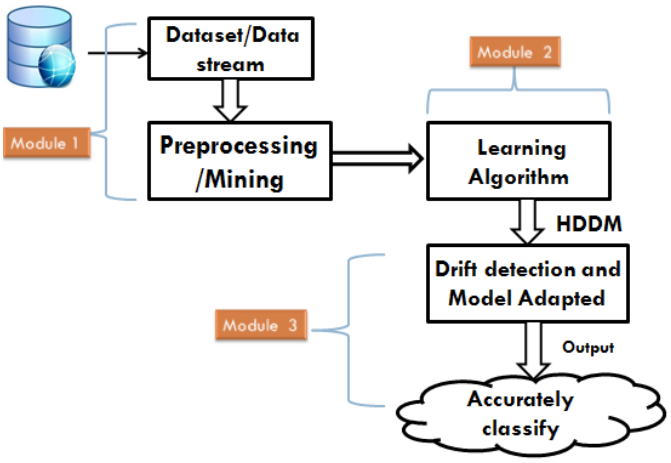

Fig- 2: System architecture

6. CONCLUSION

This paper explains about the problem of concept drift. It summarizes the necessity, types and reasons for concept change. The several concept drift detection methods viz. DDM, EDDM, ECDD, ADWIN, STEPD and DOF are discussed and methods to adopt and detect concept change. So, drift detection problem solve using one of learning algorithm and Hoeffding's Bound Algorithm. Drift detection algorithm works as per change in data is occurring. Accuracy is improved during classification because as per change occur in data, classification algorithm and drift detection algorithm is adapted. The proposed method evaluate the performance of each respective classifier in order to detect drift along with classification and improve the accuracy of classifiers by training the miss classified instances .Thus the key factor of accuracy improvement can be achieved.

REFERENCES

[1]. J Gama, P Medas, G Castillo and Pedro Rodrigues(2004), "Learning with Drift Detection", Lecture Notes in Computer Science, Vol. 3171, pp 286-295.
[2]. J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, A. Bouchachia(2014), "A Survey on Concept Drift Adaptation ", ACM Computing Surveys, Vol. 46, No. 4, Article 44.
[3]. J Gama, P Medas, G Castillo and Pedro Rodrigues(2004), "Learning with Drift Detection", Lecture Notes in Computer Science, Vol. 3171, pp 286-295.
[4]. M Baena-Garcia, J Campo-Avila, R Fidalgo, A Bifet, R Gavaldµa and R Morales-Bueno(2006), "Early Drift Detection Method", IWKDDS, pp. 77–86.
[5]. A Bifet(2009), "Adaptive Learning and Mining for Data Streams and Frequent Patterns", Doctoral Thesis.
[6]. 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.

[7]. 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.

[8]. 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.

[9]. Isvani Fr□ıas-Blanco, Jos e del Campo- Avila, Gonzalo Ramos-Jim enez, Rafael Morales-Bueno, Agust□ın Ortiz-D□ıaz, and Yail□e Caballero-Mota," Online and Non-Parametric Drift Detection Methods Based on Hoeffding's Bounds" VOL. 27, NO. 3, MARCH 2015.

[10]. Ludmila I. Kuncheva(2004), "Classifier ensembles for changing environments", Multiple Classifier Systems, Lecture Notes in Computer Science, Springer , vol. 3077, pages 1–15.

[11]. W. N. Street and Y. Kim(2001), "A streaming ensemble algorithm (SEA) for large-scale classification," in Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 377–382.

BIOGRAPHIES

Kranti wadewale received the Bachelor's degree in Computer science and Engineering from M.G.M College of Engineering, Nanded in 2014. Her main area of interest includes BIG Data Mining. She is now pursuing Masters in Computer Engineering from MIT College of Engineering, Pune.

Prof. Sharmishta Desai is working as professor in MIT College of Engineering. Her Research Interest are Information Security, Machine Learning.