A New Classifier for Handling Concept DriftingData Stream

[1]Ms. Rucha Patil, [2]Mr. S.B.chaudhary

Department of Computer Science and Engineering
JSPM's JSCOE, Pune, India.

Abstract- **Concept drifting stream data mining have recently garnered a great deal of attention for Machine Learning Researcher. The major challenges in stream data mining are focused on speed of data arrival, changes in data distribution in certain time, storage capability that uses less memory, and adapting changes in small amount of time. In this paper, a new Classifier based on hybrid approach is proposed that handle concept drifting stream data. The proposed classifier is used Naives Bayes as base learner for classification of concept drifting stream data where as concept drift is detected and handled by using drift detection method.**

Keywords: **Concept drift, stream data, classification, drift detection.**

1. INTRODUCTION

With the advent of dynamic and evolving nature of data generation environment such as the web, and other technologies has caused a fundamental change to the distribution of data such data called as Stream Data. Stream Data has distinct qualities that differentiate it from traditional data. Stream Data is now more than ever highly distributed, loosely structured, increasingly large in volume and changing over time. Broadly speaking firstly, the volume of amount of data increasing exponentially each year and secondly the speed at which the new data is being generated of distinct concept and changes over time. Stream Data is generated by a number of sources including telecommunication, social networking, radio frequency identification, scientific data, financial data and use of other data generating applications such as online purchase transactions, stock trades every day.

Classification of such data streams has become an important area of machine learning. Traditional classification techniques of machine learning assume that data have stationary distributions. Examples of such data streams applications include text mining, information filtering credit card fraud detection, email spam detection, etc.

One of the challenges in data stream classification is that the underlying distribution of data generation process of a stream tends to changes over time, called concept drift. A model learned from an earlier part of stream data loses its classification accuracy upon the arrival of new instances that exhibit concept drift. An appropriate method for such problems should adapt to drifting concepts by revising and refining the method as new data become available, without the need to store all data. Several important approaches such as single and ensemble classifier that are developed so far, handle gradual and abrupt concept drift in data stream but not accurately enough[1] [2] [3]. In addition, analyzing real word data by using such approach is very difficult and hence need more ground attention. Hybrid approach over single classifier for data streams has been proven both theoretically and experimentally. Accordingly, in this paper, a new classifier is proposed for classification of concept drifting data streams. The proposed classifier classifies stream data using naive's bayes and handle concept drift using DDM method.

The paper is organized as, Related Work discussed in Section 2. Section 3 discusses the Proposed Work, Section 4 described experimental results and finally with conclusion and future work in Section 5.

2. RELATED STUDY

A number of single and ensemble approaches are introduced in machine learning by several Researchers for concept drift handling. Here review of those method included in this paper.

In 1986, the first systems capable of handling concept drift were STAGGER [5] and then FLORA System in 1996 [6]. These approaches are used to handling concept drift. The FLORA System maintains a Dynamic Window to keep track of occurrences of Drift, but it has limitation on the speed of arriving data.
In 2000, another approach based on decision tree method such as VFDT in 2000, CVFDT in 2001, and OVFDT in 2011 proposed recently and developed so far [7], [8], and [9]. The VFDT method can process each example

in constant time and memory being able to incorporate tens of thousands of examples per second using off the shelf hardware but inability to cope with concept drifts. The CVFDT is an extended version of VFDT which handle concept drift that uses sliding window and monitor the affect of sample in sliding window on current decision tree accuracy. The OVFDT is also one of the methods in this category. A significant feature of OVFDT is its ability to reduce the decision tree size learnt from massive data streams and have better accuracy than VFDT.

In 2001, there are several approaches based on Ensemble classifier such as SEA, weighted majority and DWM seems to be an effective. The Streaming Ensemble Algorithm (SEA) copes with concept drift with a bagging ensemble of C4.5 classifiers [10]. SEA reads a fixed amount of data and uses it to create a new classifier. Performance of this method is measured over the most recent predictions based on the performance of both the ensemble and the new classifier. The Weighted Majority provide the general framework of weight processing of some fixed expert system by changing integration rule of each basic classifier[11], [12]. WM is able to track the occurrence of concept drift but cannot dynamically add and delete expert with occurrence of concept drift. Another approach in this category is Dynamic Weighted Majority (DMW) deals with data stream arriving as a single sample but it can be easily extended to handle datastream arriving as sample block. It can dynamically add and delete expert with occurrence of concept drift, and [13].

The few other approaches such as DWM with naive Bayes in 2007 present an ensemble method for concept drift that dynamically creates and removes weighted experts in response to changes in performance [13]. Other approach such as Pared Learner (PL-NB) in 2008 that uses the naives bayes as base learner. Pared Learner (PL-NB) used the online version of naive's bayes [14]. It suggests that, paired learners outperformed or performed comparably to learners more costly in time and space. In some cases, other methods required between 10 and 50 base learners to obtain high accuracy on the problems considered, but this method used two. Ironically, for one problem, this obtained the best performance for two methods when their ensembles had two members.

3. PROPOSED WORK

The proposed method uses Naives Bayes as base learner and Drift detection method for handling concept drifting data streams [16]. This method focuses on to improve performance of classification in terms of accuracy.

3.1 Naives Bayesian Classifier

The Naives Bayesian Classifier remains a popular classifier looking at its competitive performance in many research domains and its simplicity in computation that allows researchers to save a lot of computational costs. This is statistical classifier that is able to perform probabilistic reasoning under uncertainty using Bayes theorem that can relate the posterior distribution to three other probability distributions and it is written as,

$$\text{Posterior Distribution} = \frac{(\text{Prior} * \text{Likelyhood})}{\text{Evidencen}} \qquad (1)$$

Posterior Distribution P (C|DS): determined the classification according to the prior, likelihood and evidence.

The prior distribution P(C): represents that is known already, or previous analyses of classes.

The likelihood distribution P (DS|C): describes the probability of observing the data, given the class.

The evidence distribution P (DS): describes the likelihood of observing the data, averaged over all possible classes

Consider DS as a data sample consisting n features {
d1, d2, dn} and C denotes a class {c1, c2} to be predicted. Classification is determined by obtaining P(C|DS), probability for a class conditioned upon an observed data sample DS, is equal to its likelihood P(DS|C) times it probability prior to any observed data sample P(C), normalized by dividing evidence P(DS).

(2)

$$P(C|DS) = \frac{P(C) * P(DS|C)}{P(DS)}$$

Where P(C|DS) is Posterior Distribution like

{P(c1|d1,d2,….,dn) and P(c2|d1,d2,….,dn)}, The likelihood distribution is denoted as P(DS|C) like {P(d1,d2,….,dn|c1) and P(d1,d2,….,dn| c2)} and P(C) is class prior distribution like {P(c1) and P(c2)}

Since posterior is greater in the class c1 case, we predict the sample is belonging to Class c1 otherwise class c2.

However, the discussion is concern, one common rule is considering the hypothesis that is most probable, and this is known as the maximum posteriori. The corresponding classifier is defined as
For Categorical Data

$$C = \operatorname{argmax}_{Ci} P(Ci) * \prod_j (vj|Ci) \quad (3)$$

For Numerical Data, it stores the sum of an attribute's values and the sum of the squared values.

$$(vj|C) = \frac{1}{\sqrt{2\pi\sigma ij^2}} e^{-(vj-\mu ij)*2\sigma ij^2} \quad (4)$$

Where vj is the jth attributes value, μij is the average of the jth attribute's values for the ith class, and σij is their standard deviation.

3.2 Method for Concept Drift handling

There are approaches that pay attention to the number of misclassification produced by the learning model during prediction. In learning approach, the model must make a prediction when an example becomes available. Once the prediction has been made, the system can learn from the examples and incorporate it to the learning model.
The method that proposed in this paper, called Drift Detection Method (DDM) has been developed to improve the detection in presence of concept drift. The drift detection method uses a binomial distribution that distribution gives the general form of the probability for the number of error in a sample of n examples.
(5)

For each time step i in the sequence of examples, the probability of misclassifying (pi) $pi = \frac{error(i)}{i}$ is considered to be error rate, with standard deviation given by

(6) $$si = \sqrt{pi * (1 - pi)^2 / i}$$

A significant increase in error of the method means that changes in class distribution and, hence, the actual learned model is supposed to be inappropriate.
Here following conditions are to be checked:

For the warning level (pi+si >pmin+2*smin): this level indicates that drift may be occurred, after this level, the examples are stored in hope of a possible change of context.

For the drift level (pi + si >pmin + 3 *smin): this level indicates that the concept drift is supposed to be true, and once the drift is detected, at the same time there is need to reset the learning method and hence a new model is to be learn using the instances stored since the warning level.

4. EXPERIMENTAL STUDY

Equalize Propose work is implemented in java and evaluated on synthetic dataset Stagger. The result shows that using this approach accuracy is improved.

4.1 Data Stream Generation

STAGGER [6] is used for simulated concept drift, total changes in concept descriptions. Stagger data generated using three attributes color ∈ {red, green, blue}, shape ∈ {rectangular, circular, triangular}, and size ∈ {small, medium, large}. Three blocks of data are defined as follows. In the first block, an instance is labelled 0 if color = red ∧ size = small. In the second block, an instance is labelled 0 if color =green ∨ shape = circular, and in the third block if size = medium∨ large.

4.2 Error Vs Drift Level Evaluation

The figure 1 shows the error of misclassification, warning level and drift level indicated blue line, red line and green line respectively. It shows that when the error is increase the possibility of concept drift is there when the error level crosses the warning level and once the error reach or cross drift level then the drift is confirmed and hence drift is detected.

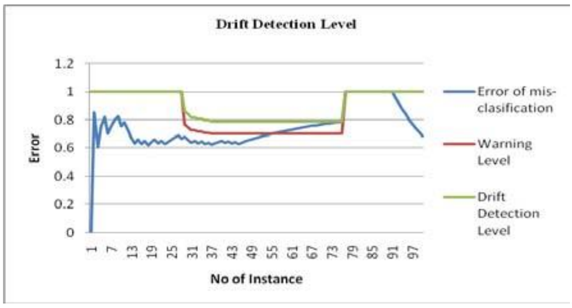

Figure 1: Show error of misclassification, warning level and drift level indicated series 1, series 2, and series 3 respectively.

4.3 Accuracy Evaluation

The figure 2 shows horizontal axis as number of instance and vertical axis as % accuracy. The blue color series 1 is the accuracy of before handling and figure 3 describe the accuracy of after handling by using drift detection method. And in figure 4 shows the comparison of accuracy of before and after drift handling.

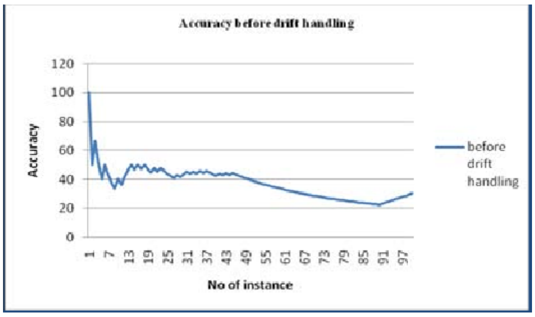

Figure 2: Show no of instances Vs Accuracy in %, before handling drift**.**

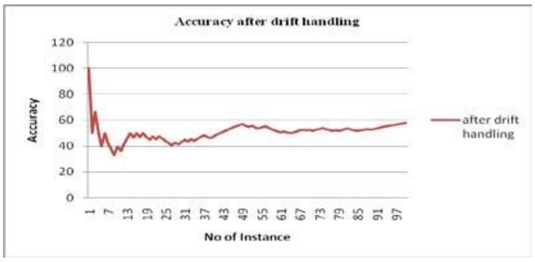

Figure 3: Show no of instances Vs Accuracy in %, after handling drift.

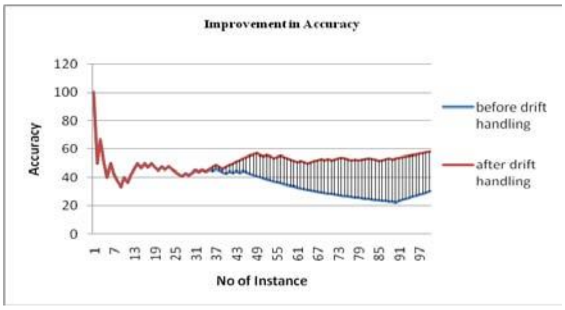

Figure 4: Show no of instances Vs Accuracy in %, series 1 is the accuracy of before handling drift and after handling drift.

It is noted and observed that proposed approach is improving the accuracy of classification.

5. CONCLUSION

Mining concept drifting data streams is a challenging research in machine learning. In particular, this paper incorporated the proposed work plan for classification of data stream in presence of concept drift. The naives bays is used as base learner and for handling concept drift, prediction result then submitted to the drift detection method for checking drift level, once the drift is detected, a new model islearnt using the examples stored since the warning leveltriggered. Hence by this way, the concept drift is handled.

REFERENCES:

1. Jeonghoon Lee, Frederic Magoules, “Detection of Concept Drift for Learning from Stream Data” IEEE 14th International Conference on High Performance Computing and Communications, 2012
2. OUYANG Zhenzhen, “Study on the Classification of Data Streams with concept Drift”, Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011
3. Mahnoosh Kholghi, Hamed Hassanzadeh, Mohammad Reza Keyvanpour, “Classification and Evaluation of Data Mining Techniques for Data Stream Requirements”, International Symposium on Computer,Communication Control and Automatio, 2010.
4. C. Agrawal, J. Han, J. Wang, P. Yu, “A Framework for On-Demand Classification of Evolving Data Streams”, IEEE Transactions on Knowledge and Data Engineering, Volume 18(5), pp 577-589, 2006.
5. J. C. Schlimmer and R. H. Granger, “Beyond incremental processing: Tracking concept drift”, In Proceedings of the Fifth National Conference on Artificial Intelligence, pages 502–507. , AAAI Press, Menlo Park, CA, 1986.
6. G. Widmer and M. Kubat, “Learning in the presence of concept drift and hidden contexts”. In Machine Learning, 1996, 23 -69-1 0 I.1996
7. P. Domingos and G. Hulten, "Mining high-speed data streams", In Knowlwdge Discovery and Data Mining, 2000, pp. 71-80.
8. G. Hulten, L. Spencer and P. Domingos, "Mining time- changing data streams", In Proc. ACM SIGKDD, San Francisco, CA, USA, 2001, pp.97-106
9. Hang Yang, Simon Fong, "OVFDT with Functional Tree Leaf -Majority Class, Naive Bayes and Adaptive Hybrid Integrations, 3rd International Conference on Data Mining and Intelligent Information Technology Applications (ICMiA), 2011
10. W. Street and Y. Kim. A streaming ensemble algorithm (sea) for largescale classification. In int'IConf. on Knowledge Discovery and Data Mining (SIGKDD),2001.
11. Blum. Empirical support for winnow and weighted majority algorithms: Results on a calendar scheduling domain. Machine Learning, 26:5–23, 1997.
12. N. Littlestone and M. K. Warmuth, The weighted majority algorithm. Information and Computation, 108:212–261, 1994.
13. J Z Kolter and Marcus A. Maloof. “DynamicWeighted Majority: An Ensemble Method for Drifting Concepts”.Journal of Machine Learning Research 8 (2007) 2755- 2790, 2007
14. Stephen H. Bach and Marcus A. Maloof, “Paired Learners for Concept Drift”, Eighth IEEE International Conference on Data Mining, 2008.
15. Joao Gama, Pedro Medas, Gladys Castillo, and Pedro Rodrigues “Learning with Drift Detection” Lecture Notes in Computer Science, v. 3171 Springer Verlag, 2004 pp. 286.