Proc. of the Second Intl. Conf. on Advances in Electronics, Electrical and Computer Engineering -- EEC 2013

ISBN: 978-981-07-6935-2

Classification of Stream Data in the presence of Drifting Concept

Sudhir Ramrao Rangari
Ms. Snehlata S Dongre
Dr. L G Malik

***Abstract*— Mining stream data have recently garnered a great deal of attention for Machine Learning Researcher. The major challenges in stream data mining are drifting concept that deals with data whose nature changes over time. Concept drift is one of the core problems in Stream data mining and machine learning. Classification of Stream data in the presence of drifting concepts is more difficult and one of the core issue. In this paper, A Classifier based on hybrid approach is proposed and implemented 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. Experiments and results on datasets show that the proposed approach performs well with improvement in accuracy of classification and can detect and adapt to concept drifts.**

Keywords—concept drift, stream data, classification, drift detection.

I. Introduction

The problem of data stream classification has been widely studied over last decade. The dynamic and evolving nature of data stream gain attention towards research in the field of data stream mining. Classification of such data streams has become an important area of machine learning. Traditional classification techniques of data mining and machine learning assume that data have stationary distribution. Examples of such data streams applications include text mining, information filtering credit card fraud detection, email spam detection, etc 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.

Mr. Sudhir Ramrao Rangari (M Tech CSE)
Department of Computer Science and Engineering
G. H. Raisoni College of Engineering
Nagpur, India

Ms. Snehlata S. Dongre
Department of Computer Science and Engineering
G. H. Raisoni College of Engineering
Nagpur, India

Dr Latesh Malik
Department of Computer Science and Engineering
G. H. Raisoni College of Engineering
Nagpur, India

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 with distinct concept and changes over time. Stream Data is generated by a number of sources including telecommunication, social networking, scientific data, credit card data and use of other data generating applications such as online purchase transactions, stock trades every day.

Dealing with data whose nature changes over time is one of the core problems in data mining and machine learning. The major challenge in data stream classification is that the underlying distribution of data generation process of a stream tends to changes over time. The underlying data distribution may change and these changes make the model built on old data inconsistent with the new data and regular updating of the model is necessary. This problem is popularly known as drifting concept in data distribution.

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]. Analyzing real word data by using such approach is very difficult and expensive 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 Stream data in the presence of drifting concept. The proposed classifier classifies stream data using naive's bayes and handle concept drift using Drift Detection Method in order to improve the accuracy of classification.

The paper is organized as, Related Work discussed in Section II. Section III discusses the Proposed Work, experimental setup and result in IV and finally with conclusion in Section V.

II. Related Work

The first systems capable of handling concept drift were STAGGER [5] and FLORA [6] System. 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.

Another approach based on decision tree method such as VFDT [7], CVFDT [8], and OVFDT [9] proposed recently and developed so far. The VFDT method can process each

Proc. of the Second Intl. Conf. on Advances in Electronics, Electrical and Computer Engineering -- EEC 2013

ISBN: 978-981-07-6935-2

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.

On other hands, there are several approaches based on Ensemble classifier such as SEA [10], weighted majority [11] [12] and DWM [13] seems to be an effective. The Streaming Ensemble Algorithm (SEA) copes with concept drift with a bagging ensemble of C4.5 classifiers. 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. 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 data stream arriving as sample block. It can dynamically add and delete expert with occurrence of concept drift.

III. Propose Work

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

A. 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,

$$\text{Posterior Distribution} = \frac{(\text{Prior} * \text{Likelyhood})}{\text{Evidences}} \quad (1)$$

Consider DS as a data sample consisting of 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).

$$\mathrm{P(C|DS)} = \frac{\mathrm{P(C)} * \mathrm{P(DS|C)}}{\mathrm{P(DS)}} \quad (2)$$

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

$$\mathrm{C} = \mathrm{argmax_{Ci}}\, \mathrm{P(Ci)} * \prod_{\mathrm{j}} \mathrm{(vj|Ci)} \quad (3)$$

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

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

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.

B. Drift Detection Method

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

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.

For each time step instance in the sequence of examples, error is defined as the number of misclassification

error=number of misclassification

The probability of misclassifying (pi) is considered to be error rate,

$$Pi = \frac{error}{i} \quad (5)$$

with standard deviation given by

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

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.

Proc. of the Second Intl. Conf. on Advances in Electronics, Electrical and Computer Engineering -- EEC 2013

ISBN: 978-981-07-6935-2

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.

IV. Experimental Setup and Result

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

A. Data Set

STAGGER [6] is used to simulate 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}. There three target concept that are generated randomly, the first block define the concept 1 and it is labeled 0 if (color = red ∧ size = small). The second block define the concept 2 and it is labeled 0 if (color =green ∨ shape = circular), and the third block defined the concept 3 and it is labeled 0 if (size = mediumV large). The target concept other than three is labeled 1.

TABLE 1. TARGET CONCEPT AND CLASS LABEL

Color	Shape	Size	Class	Target Concept
red	rectangular	**small**	0	Concept 1
green	triangular	small	0	Concept 2
Blue	**circular**	small	0	Concept 2
Red	rectangular	**medium**	0	Concept 3
Blue	triangular	**large**	0	Concept 3
Blue	triangular	small	1	otherwise

The data is generated using above concept and drift is induced in data distribution to perform the evaluation.

B. Result

The generated data is supplied for training and build the model and make it as current model. The Testing instance is given for classification according to existing model prediction is made. The result of the prediction is then passing to detect the concept drift. If the concept drift is detected then the model is rebuild with the support of instances encounter since warning level to drift level.

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 the significant increased in error indicate that change in distribution, when error crosses the warning level, there is possibility of concept drift is and once the error reach or cross drift level then the drift is supposed to be true and hence drift is detected.

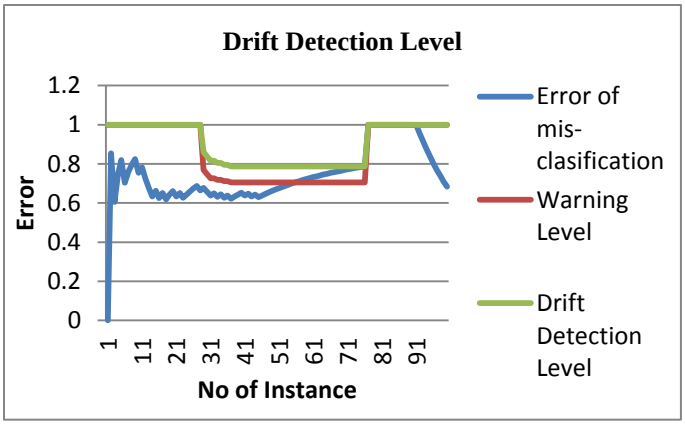

Figure 1: shows error of misclassification, warning level and drift level indicated blue, red and green respectively.

The figure 2 shows horizontal axis as number of instance and vertical axis as % accuracy. The series 1 is the accuracy of before handling and series 2 is the accuracy of after handling by using drift detection method.

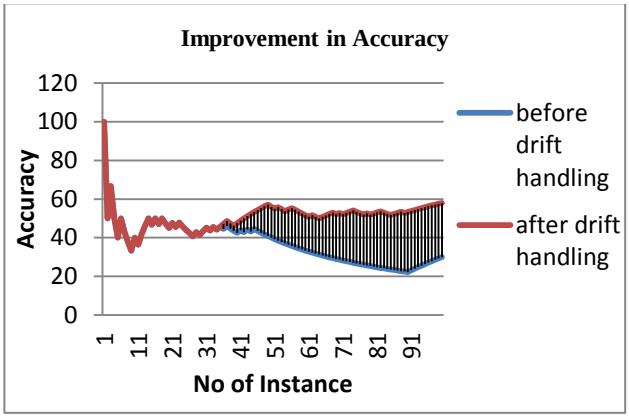

Figure 2: Show no of instances Vs Accuracy in %, series 1 is the accuracy of before handling drift and after handling drift

It is observed that, proposed approach is improving the accuracy of classification

The below table 2 shows the analysis on total number training and testing instances with correctly classified instances and accuracy.

TABLE 2. ACCURACY OF CLASSIFICATION OF DATA BLOCK

Total Instance	Training Instance	Testing Instance	Data Block	Correctly Classified	Accuracy (%)
2000	500	1500	500	464	92.80
			1000	937	93.98
			1500	1387	92.65

ISBN: 978-981-07-6935-2

The experiment performs on 2000 instances, which are divided into training and testing set. The first 500 instance are used for training and remaining instances are used for testing and it is divided into three blocks with different concept.

v. Conclusion

Data streams mining in the presence of concept drift is a challenging research in machine learning. The Bayesian classifier is used as base learner and prediction result of testing then submitted to the drift detection method for checking drift level, once the drift is detected, a new model is learnt using the examples stored since the warning level triggered. Hence by this way, the concept drift is handled. The result of using this approach is improving the accuracy of classification.

References

[1] Jeonghoon Lee, Fr´ed´eric Magoul`es, "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] A. 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 M. A. Maloof. Dynamic weighted majority An ensemble method for drifting concepts.

[14] J. Zico Kolter and Marcus A. Maloof. "DynamicWeighted Majority: An Ensemble Method for Drifting Concepts". Journal of Machine Learning Research 8 (2007) 2755-2790, 2007

[15] Stephen H. Bach and Marcus A. Maloof, "Paired Learners for Concept Drift" Eighth IEEE International Conference on Data Mining, 2008.

[16] 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,