Handling Concept Drift in Data Stream Classification.

Ritika Jani, Nirav Bhatt, Chandni Shah

Abstract: Data Streams are having huge volume and it can-not be stored permanently in the memory for processing. In this paper we would be mainly focusing on issues in data stream, the major factors which are affecting the accuracy of classifier like imbalance class and Concept Drift. The drift in Data Stream mining refers to the change in data. Such as Class imbalance problem notifies that the samples are in the classes are not equal. In our research work we are trying to identify the change (Drift) in data, we are trying to detect Imbalance class and noise from changed data. And According to the type of drift we are applying the algorithms and trying to make the stream more balance and noise free to improve classifier's accuracy.

Index Terms: Data Stream mining, classification, semi supervised learning, concept drift.

I. INTRODUCTION

Data streams are having continuous, unwoundable flow of data, so there are chances of misclassifying the samples from another class. The problem of imbalance class occurs when there are more number of samples in one class than the other class.[1] With this problem the model would not able to have the correct accurate prediction. At the other side concept-drift in streams is when the data is changed and the model is not able to predict the change. These issues in data stream may cause the real world problems like credit card fraud, facial recognition etc. And misclassification of class may add noise in data, and due to noise in data the primary concept may change. Many of the time the classifier is rebuts to the noise [10]. Moreover the change in the data does not mean the change in the hidden context of the data [11]. The fundamental rule of getting the information is to remove the noise [12]. There are basically two types of approaches in concept drift. General framework of concept drift is given in [9] for data stream classification. First one is single learning approach. The second one is ensemble learning approach which is having STAGGER, FLORA, ACE, SEA algorithms[13][14].

One Real time application where the Class imbalance and concept drift problem accrue is the Credit card fraud transections, where all the transactions are going to be the normal one or the genuine one. But if there were any fraud transection accrue, it is difficult to major the fraudant transection because most of the samples are biased towards the majority classes or we can say to the normal transections. If there is any change of the amount accrue in the transection we can not identify if it is because of drift or because of the imbalance class.

Revised Manuscript Received on August 05, 2019.

Ritika Jani has obtain her master degree in Information Technology from Chandubhai S Patel Institute of Technology, CHARUSAT University, India.

Nirav Bhatt is working at Department of Information Technology in Chandubhai S Patel Institute of Technology, CHARUSAT.

Chandni Shah is working at Department of Information Technology in Chandubhai S Patel Institute of Technology, CHARUSAT.

II. PROPOSED SYSTEM

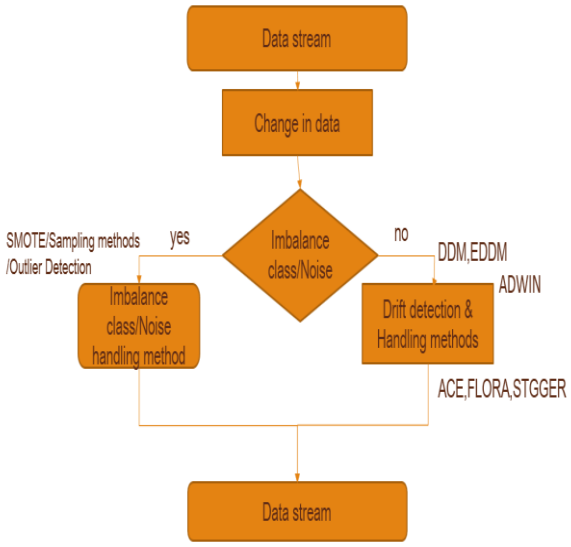

Fig 1: Proposed system

A. Flow Steps of the system:

The proposed system describes the work to be carried out in order to achieve the objective of the research work. Firstly Detail data stream mining are studied in detail. An important information about Concept Drift in data stream are focused, also the algorithms for particular type of drift has been studied. The factor which is reason behind changing nature of the data is somehow the imbalance class stream and noise.

Data Pre-Processing.

When we have a continuous Unbounded Data, we divided it in chunks of the stream for ease of access. Pre-processing of the stream must require in terms of missing values and outliers.

Step 1: Continuous Unbounded Data to Blocks of data.
Input – Unbounded size of the stream, which can't be stored in memory
Output- Divided chunks of the stream for ease of access.

Step 2: Detecting the change from data
Input – Block of the data stream.
Output – Change in data streams.

Action – as soon as we find any change in data we have to find whether it is by imbalance class or not.

Step 3: Correction of data (if data has changed because of the imbalance class)
Input – Changed Data Stream.
Output – Balance class form the imbalance data stream and divide the stream into testing and training dataset to check the accuracy of prediction.
Action – If the data has been changed by noise then it might be possible that few of the samples are misclassified or we can say noise may produce by imbalance stream. Pre-processing data which has no missing values and outliers.

Step 4: Finding Drift
Input – Continuous Data Stream
Output – Detected Drift
Action: Detected Drift

End

B. Handling Imbalance class and Noise.

Generally imbalance class means the number of sample in one class is more than the number of samples in another class. There are many methods via which we can handle the imbalance class problem. The popular methods are under-sampling and oversampling and the popular algorithms are SMOTE, SVM [7][8]. As soon as the data has been changed we check whether it is by imbalance class or not. If the data has been changed because of imbalance class than obtain Balance class form the imbalance data stream using sampling techniques and divide the stream into testing and training dataset to check the accuracy of prediction.
If the data has been changed by noise then it might be possible that few of the samples are misclassified or we can say noise may produce by imbalance stream.
To check any change in data stream in term of amount in credit card transaction, we have applied drift detection methods, and then we have applied drift solving methods according to the type of a drift.

III. EXPERIMENT DETAILS

For Experiments we have used the Anaconda3 Cloud along with Python3.6 on windows OS with core i5-3210M @2.50 GHz processor with 4 GB RAM. We have used the Credit card dataset to find the imbalance class and Drift. We have noticed that most of the transactions are normal one. There are few transection of fraudulent transection but due to their minority against the majority classes (normal classes) all the incoming samples are going to be bias towards the majority once, and it is difficult to predict that the incoming transection is going to be ideal transection.
First task is to identify and remove or fill the missing values in dataset. Then we have applied the sampling method with classification algorithms to train and test the classifier.

```
Type "copyright", "credits" or "license()" for more information.
>>> ================================ RESTART ================================
>>>
          V1        V2        V3        V4        V5        V6        V7  \
0.05 -2.899147 -1.971975 -2.389740 -2.195683 -1.702021 -1.406757 -1.434423
0.95  2.081223  1.808585  2.062635  2.566501  2.098960  3.160382  1.407632

          V8        V9       V10  ...         V21       V22       V23  \
0.05 -0.842147 -1.758426 -1.338636  ...   -0.504674 -1.081892 -0.472246
0.95  1.049984  1.780783  1.548557  ...    0.537868  1.128987  0.488016

         V24       V25       V26       V27       V28  Amount  Class
0.05 -1.143662 -0.825026 -0.697348 -0.415246 -0.317843    0.92    0.0
0.95  0.866358  0.760699  0.920915  0.387746  0.256090  365.00    0.0
```

Fig 2: Dataset overview

As we discussed in above section there are methods to solve the imbalance class problem. We use the under sampling technique to archive the balance class results.
We have resampled the random samples from the data and according to that ratio we have train and test the data with the help of python libraries and using SVM algorithm [4].

We have tested our system on credit card dataset which is having nearly more than 2 lakhs of rows. For checking the fraud transection the column labeled as class '0' and '1'. If the class is '0' than there is a normal transaction, and if it is labeled as '1' than there is a fraudulent transaction.

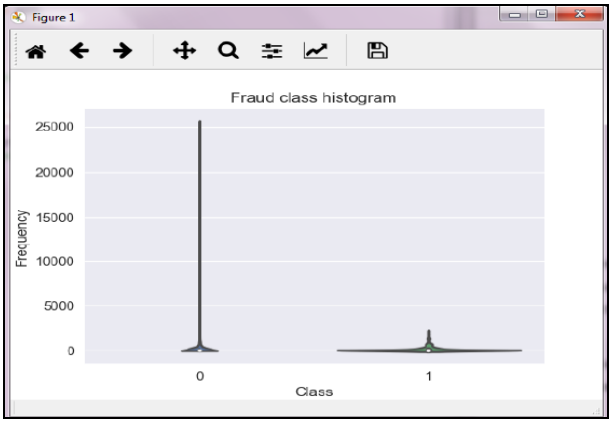

Fig 3: Plotting of labels

There are lakhs of transactions are running all the day so and very less of them are going to be the fraud one so we can say that finding fraud transactions is tough task. For that balancing the class is must require to have the better performance of the algorithm. For balancing the class we have used the under sampling method which is works best when the data is larger in the size. We have selected some random samples from the class labeled as '0' (majority class). For checking the accuracy we have split it into random training and testing data using the cross validation. After that generating the confusion matrix for checking that how many of the false positive has been detected by the classifier. The recall score is quite high for testing data (unseen data).
After that we have checked the performance of the whole classifier with ROC curve which is ultimately True Positive Rate v/s false Positive Rate. The curve is >90% accurate.

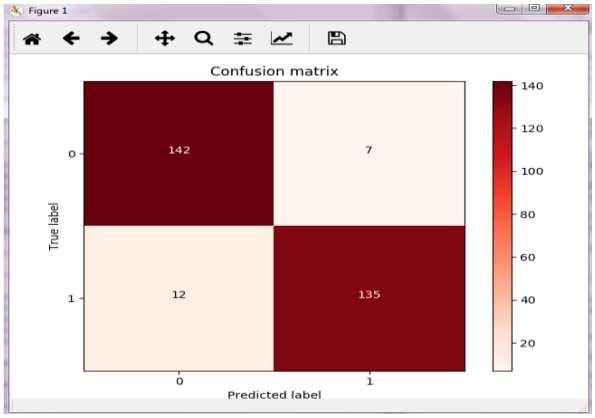

Fig 4: Confusion matrix of re-sampled dataset

The distribution generating the items of a data stream can change over time which is concept drift [9]. if the amount is high for the class labeled as '0' than that would be the fraud and it isn't, it is a drift for a class labeled as '1'. DDM (Drift Detection Method) is one of the method which counts the number of errors produced by the learner during the prediction. This method uses binomial distribution and that distribution gives the general form of the probability for the number of variables and number of errors in a sample of "n". For each i^{th} point which is sampled in sequence (s_i), the rate of error is misclassifying probability (p_i) [9]. The standard deviation is given by the equation as below.

$$s_i = \sqrt{pi(1 - pi)}/i \quad (1)$$

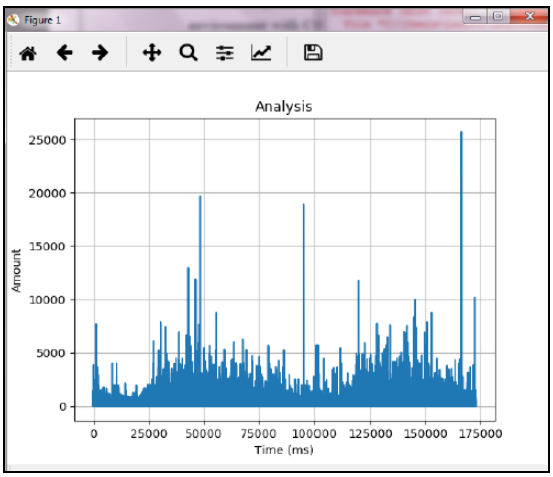

Fig 5: Analysis of Drift Deteection Method

There are several methods to deal with the drift, few of them like decision tree methods, Windowing methods etc. We have used the naïve Bayes and the adaptive windowing technique to overcome the drift. If the window size is smaller the results are being more accurate.

IV. CONCLUSION

One of the main issue with data stream mining is the Concept Drift, So identifying the drift and noise is the tough task. If some of the samples from the class misclassified (classifier is imbalanced) and data has been changed. Due to that misclassification noise may accrue. If some changes may accrue in dataset than we can't say directly that it is a true drift. When we clean up the dataset in terms of imbalance class and noise and after applying the drift detection methods then the accuracy of the classifier is overall going to increase

V. FUTURE WORK

The proposed work can be expended with some tools like R and MOA with the bigger size of dataset.

REFERENCES

1. Lara Lusa, Rok Blagus,"The class-imbalance problem for high-dimentional class prediction", 11th International Conference on Machine Learning and Application.2012
2. Joao Gama, Indre Zliobaite,Albert Bifet,Mykola Pechenizky,Abdelhamid Bouchachia,"A syrvey on Concept Drift Adaption",ACM,2014.
3. Radhika Kotecha, Sanjay Garg,"Data Streams and Privacy: Two Emeging Issues in Data Classification", 5th Nirma University International Conference on Engineering, 2015.
4. Fabricio Breve, Liang Zhao, "practical Competition and Cooperation in Networks for Semi-Supervised Learning with Conceept Drift", World Congress on Computational intelligence, IEEE, 2012.
5. Rukshan Batuwita, Vasile Palade,"Class Imbalance Learning Methods for Support Vector Machines" Singapore-MIT Alliance for Research and Technology Centre; University of Oxford, 2012.
6. Archana Purwar, Sandeep Kumar Singh,"Issuess in Data mining: A comprehensive survey", International Conference on Coputational Intelligence and Computing research, IEEE, 2014.
7. Rukshan Batuwita, Vasile Palade, "Class Imbalance Learning Methods for Support Vector Machines", Singapore-MIT Alliance for Research and Technology Centre; University of Oxford, 2012
8. Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, Guangtong Zhou,"On the Class Imbalace Problem", Fourth International Conference on Natural Computation,IEEE,2008.
9. Manuel Baena-Garcia, Jose del Campo-Avila, Raul Fidalgo, Albeert Bifet, Richard Gavalda , Rafael Morales-Bueno,"Early Drift Detection Method", Spain
10. Xiaoye WANG, Bingjie CHEN, Fie CHANG, "A Classification Algorithm for Noisy Data Streams with Concept-Drifting", Journal of Computational Information Systems, 2011.
11. Petr Kosina,"Data Stream Mining and Concept Drift: Novel Approaches and Applications". 2010.
12. Rok Blagus, Lara Lusa, Evaluation of SMOTE for high-dimensional class-imbalanced microarray data. IEEE, 2012.
13. T.Ryan Hoens, Robi Polikar, Nitesh V. Chawla,"Learning from streaming data with concept drift and imbalance: an overwiew", Springer, 2012.
14. Pedro Domingoes, Geoff Hulten,"Catching up with the data: Research Issues in Mining Data Stream", University of Washington.

ORGANIZERS

Ritika Jani has obtain her master degree in Information Technology from Chandubhai S Patel Institute of Technology, CHARUSAT University, India. She is currently working as assistant professor in computer engineering department in Dr. Jivaraj Maheta Institute of Technology. Her research area is in data stream mining & information retrieval.

Nirav Bhatt is working at Department of Information Technology in Chandubhai S Patel Institute of Technology, CHARUSAT. He had received degree of Master of Engineering in Computer Engineering from Dharmsinh Desai Institute of Technology and currently pursuing his Ph.D. in the area of Big Data Stream Analytics. His research interests include Database System, Data Mining and Big Data Stream Analytics. He is also a member of Computer Society of India and ACM and member of ACM Chapter at the institute. He is also coordinator of SWAYAM-NPTEL Local Chapter which is the National MOOCs portal being developed by MHRD, Govt. of India.

Chandni Shah is working at Department of Information Technology in Chandubhai S Patel Institute of Technology, CHARUSAT. She had received degree of Master of Engineering in Information Technology. Her area of interest is operating system, parallel computing, data mining and distributed system.