RESEARCH ARTICLE OPEN ACCESS

Concept Drift Detection Approach Towards Imbalanced Data Stream

Aishwarya M Y*, Dr Minavathi**

**Student, Department of Information Science & Engineering, PES College of Engineering, Mandya-571401*
*** Professor & Head, Department of Information Science & Engineering, PES College of Engineering, Mandya-571401*

ABSTRACT
Data stream has grown very rapidly and it is present in many real time applications. Due to its huge data which arrive with high rate, it is difficult to extract and visualize this data. Non-stationary data streams usually are affected by the phenomenon of concept drift. Detecting this drift and adapting to it becomes a challenging task while classifying data streams. Imbalanced data is another issue of data mining in the field of data stream. Imbalanced class is one of the extensive characteristics which need to be handled to improve the accuracy of the classifiers. Hence, in this paper a Concept Drift Detection Approach (CDDA) towards imbalanced data is presented. The CDDA detects the concept drift using the Jaccard distance and keeps the classifier up-to-date. The study has been done on artificial data set as well as real data set. The finding of this study shows good performance on the majority class and in reacting to gradual and sudden drift in comparison with the state-of-the-art models.

Keywords: Data Stream, Concept Drift, Imbalanced Data, Jaccard Similarity.

Date of Submission: 09-10-2023 Date of acceptance: 22-10-2023

I. INTRODUCTION

In the last decade generating automatic data and storing them have become quicker than ever, due to the vast progress in both software and hardware technology. Such information which is collected from these sources is called data stream. Data stream has grown very rapidly and it is present in many real time applications. Due to its huge data which arrive with high rate, it is difficult to extract and visualize this data. Non-stationary data streams usually are affected by the phenomenon of concept drift. A drift undetected leads to the drastic drop in the classifier's accuracy. Detecting this drift and adapting to it becomes a challenging task while classifying data streams. In the recent literature substantial study and analysis has been done in the field of data stream whose main objective is to handle the difficulties of stream mining with the concept drift.

Imbalanced data is another issue of data mining in the field of data stream. Imbalanced class is one of the extensive characteristics in the machine learning which needs to be handled to improve the accuracy of the classifiers. Skewed data leads to bias on the majority class label and it is very hard to predict when the minority class will be correctly classified. Hence, in this paper a Concept Drift Detection Approach towards imbalanced data is presented. The CDDA work on binary classifier handles the imbalance class using the under-sampling technique without losing the information. The CDDA detects the concept drift using the Jaccard distance and keeps the classifier up-to-date. The study has been done on artificial data set as well as real data set. The finding of this study shows good performance on the majority class and in reacting to gradual and sudden drift in comparison with the state-of-the-art models.

II. DATA STREAMING

Data generation rate is exponentially moving up due to the increase of data usage by both offline/online users, and that is why data stream comes into picture. Data streaming is defined as the data being produced which is fast-changing, temporally ordered, massive size, and continuous. Many applications in real life are continuously generating data stream like credit card fraud detection, network monitoring and web click stream [1]. Because of the high rate of speed of the data stream, it is not practical to scan the data more than once, which is denoted as one-pass constraint [2]. Furthermore, traditional algorithm cannot be applied for data stream as it is contradicting to the one-pass constraint. Another issue which is considered as the most extensive and undesirable issues with respect to classification is the concept drift. Concept drift occurs when the distribution of data is changed over time. In the absence of the drift it is hard to maintain the balance between the most recent concept and the previous concept, which leads to form the stability-plasticity dilemma [3]. Therefore, the machine learning algorithm should identify the drift and the classifier must be updated to learn from the new concept and the old concept will be degraded so the model will be up-to-date. In most of the machine learning algorithms it is assumed that the class numbers are roughly balanced. However, class imbalance exists in many real-life applications as credit card fraud and network intrusion detection. The situation in which one class is over dominating the other class is referred to as “Skewed data”. In skewed data it is hard to tell when the prediction of the minority class is correctly predicted as the decision boundary of the underrepresented class will be inclined. Therefore, the accuracy of the minority class will be degraded inevitably [4].

III. LITERATURE SURVEY

Many algorithms have been proposed to deal with the voluminous of data, one of these algorithms is VFDT [5] which is based on Hoeffding tree to handle the new incoming instances and split the tree. Learn++ [6] suggested an ensemble classifier for each coming chunk and then joining them to make the decision on the test data. Regardless of the success that has been accomplished towards learning in the existence of the concept drift, a couple of models have been proposed to take the imbalanced data so far. Generally speaking, learning in skewed data refers to the domains in which certain types of examples are underrepresented compared to the other class. Many algorithms exist to tackle the imbalanced data [7], [8] but these models contradict the constraints of data stream as these models take all data into the memory so it cannot be applied to handle the imbalanced data in the streaming data.

In the present context of the skewed data in streaming data, a significant classifier is in need to handle the imbalanced data as well as handling the concept drift which are not effectively addressed by the several previously proposed models. Keeping this in view, in this paper, the following issues are addressed: (i) Make a balance on the accuracy of the majority and minority observations so that the underrepresented class will be intensified without sacrificing the accuracy on the majority class. (ii) Classifier incremental version to update the model so that the outdated concept will be removed and the new concept will be captured when the distribution is changed. Hence, the key contribution of this paper can be summarized as follows: The skewed data is tackled by our model CDDA in the initial training set by applying the under-sampling technique. Therefore, the decision boundary of the classifier will not inline towards the minority class and the performance will be improved without scarifying on performance of the minority class.

IV. CONCEPT DRIFT DETECTION

To detect the drift in the data stream we should monitor the distribution of data whether it is stable or changed. Drift detection method is an algorithm which produces an alarm when the distribution is changed. To monitor the data and detect the concept drift the decision of the classifier is anatomized. Many approaches have been used to monitor the output of the classifier [9] and if the classification error is increased then the concept drift has occurred. In the other algorithm [10] depending on the distance of the error and the next error if it is decreased then the drift is detected. But these models are used to handle one type of the drift, it cannot react to different types of drift. In this study the model is designed to monitor the distribution of data and detect the drift and reacting to different types of drift. To achieve this objective, in this paper, a model is designed to monitor the momentum similarity of the projected values of each field of the training set and window using Jaccard similarity.

4.1 Jaccard Similarity

Jaccard similarity is one of the significant features of this study. Jaccard similarity is used to give the proximity of two sets and determine if they are shared or not. Jaccard index is used to compare similarity, distance of the sets. Measuring the similarity between sets is the result of division between the sets that are common to all divided by the union of both sets as shown in Equation 1.

$$J(V_1, V_2)=\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|} \quad (1)$$

Jaccard distances is used to find the dissimilarity between sets. It can be established by subtracting the Jaccard similarity by 1 as shown in Equation 2

$$J(V_1, V_2)=1-\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|} \quad (2)$$

In Equation (1), V_1, V_2 represents the values obtained from each field of the window as well as the values of the field which is selected from the set of trained classifiers accordingly.

4.2 Proposed Algorithm:

The algorithm for the detection of the drift in the data stream is given in the following steps having these acronyms - $W(f_i)$ - Values found in the respective field of the window;

$C(f_i)$ - Values found in the respective field of the chunk;

Sim - Similarity Ratio

1. Initialize the window from the streaming data.
2. Get all values as sets $W(f_i)$, $C(f_i)$ that is found for attributes in respective record sets of window W and chunk of the ensemble classifier C.
3. Find the distance between the two sets $W(f_i)$, $C(f_i)$ using Jaccard.
4. Aggregate the distance between the field values of the window and chunk to the value $sim_{w(f_i)\leftrightarrow c(f_i)}$
5. Calculate the average of similarity. If the similarity $sim_{w(f_i)\leftrightarrow c(f_i)}$ is greater than the threshold value then the drift has occurred and the old records should be removed from the training set and then the model should be updated accordingly.

V. EXPERIMENTAL SETUP

The CDDA performance as well as the contemporary models are tested using the synthetic and real datasets and were carried out in Java as part of the Massive Online Analysis (MOA) framework [11]. We have conducted three types of experiments. The first experiment used the synthetic dataset to show how CDDA reacts to sudden drift and how it can recover quickly from the drift. In the second experiment we have taken another synthetic dataset to show the behavior of the gradual drift and how it is detected and applying the incremental classifier accordingly. In the last experiment a real dataset was used to show the performance of our model. Different state-of-the art algorithms are used to prove the ability of our model to learn under several types of concept drift from the skewed data. All the tested algorithms are already a part and are freely available data in the MOA tool. The different experiments were conducted on PC 3.0 GHz core i5 with 8GB of memory, running Microsoft Windows 7. The different algorithms which have been used to compare with the proposed model are SERA[12], Accuracy Weighted Ensembles AWE [13], SMOTE [14], OOB [15].

5.1 Datasets Description

This subsection presents the datasets used in our experiments designed to compare the performances of our approach with other methods. Most of the common benchmarking dataset used to investigate the algorithm of the traditional classification do not contains any kind of change drift. There are two types of dataset real and synthetic dataset. In case of real dataset, it is hard to determine when the drift will start to occur, the position of the drift, or even if it is a drift or noise when we are using real datasets. Therefore, using only real datasets is hard to evaluate the performance of algorithms if the concept drift has occurred. Hence for this purpose, the streaming data are investigated using synthetic dataset. Artificial dataset is very common as we can determine the position, number and duration of the concept drift. Hence, our proposed approach is tested using 3 datasets, 2 of which are synthetic datasets and the last one is real dataset. The synthetic datasets were generated using MOA (Massive Online Analysis) framework [11] which contains artificial stream generators.

5.2 Evaluation Metrics

In case of balanced learning scenarios different evaluation metrics can be used to evaluate the performance of the classifier such as accuracy which determine the overall recognition performances of the classifier on the tested samples, and classification error which describes the total number of observations that were incorrectly classified. Accuracy and classification error were mostly identified by the majority class, and thus metrics are not appropriate to evaluate the performance in the scenario of imbalanced learning. In the imbalanced data different metrics have been put forward to evaluate the performance of the classifier. First, G-means value which is considered as significant and common metric to evaluate and analyze the imbalanced data. The high value of G-means indicates that the performance of both classes of imbalanced data is performed equally. Second, we include the metric which combines the precision and recall into one metric which is called F-measure for evaluating the performance of a learner on the minority observations. The different evaluation parameters applied in our study are shown as follows:

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F-measure:

$$F - measure = \frac{2.\ \ Recall\ .\ Precision}{Recall +\ \ Precision} \tag{6}$$

G-mean:

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{7}$$

VI. RESULTS AND DISCUSSION

In this section, we compare CDDA with three state-of-the art algorithms on artificial data sets as well as real dataset. We present the experimental results regarding the performance evaluation on different metrics like F-measures, G-mean and Accuracy.

Before we analyze the performance and compare our algorithm with other algorithms in details, we shall begin by providing the considerable observations obtained from the performance of the algorithms on all its datasets. The average values of accuracy, F-measure and G-mean and its rank for each dataset are shown in Tables 2 to 4. A summary of mean ranks of comparative algorithms on all datasets taking in consideration on all evaluation metrics is given in Table 5.

From the Tables 2 to 4 we can recognize several observations. First, in terms of accuracy AWE outperform other comparative algorithms and obtains the first mean, which means that the AWE maintains a good prediction on majority class. CDDA has got the second rank which maintains good performance on majority. However, the superior performance of AWE in terms of accuracy is causing an inferior performance on the minority class metrics as F-measure and G-mean and this is most probably due to AWE lacking the mechanism to handle the imbalance class. OOB obtains a good mean rank for imbalance class measures F-measures and G-mean and this boosts the performance on majority class which is coming at the expense of the accuracy which is having a mean rank 4.7. It can be recognized that the SMOTE, SERA, and AWE give an inferior prediction result as compared to OOB and CDDA. Finally, we can recognize that CDDA outperforms the other methods in the parameters that are used to measure the imbalanced class as F-measure, G-mean and AUE. Only on the G-mean parameter the OOB got better performance especially on big size dataset on SEA dataset and this is because the OOB oversamples the majority class. In terms of the accuracy CDDA has got the second mean compared to the other methods. We can conclude that CDDA method achieves a steady increase on performance on the minority class, without losing the accuracy of the majority class.

Table 1: Classification accuracies of different algorithms (%)

Dataset ↓/ Algorithm→	AWE	SERA	SMOTE	OOB	CDDA
Hyperplane (Synthetic dataset)	95.56(1)	92.66(3)	90.75(5)	91.94(4)	94.45(2)
SEA$_s$ (Synthetic dataset)	94.97(1)	90.97(3)	90.01(4)	88.60(5)	91.2(2)
Electricity (Real Dataset)	87.912(2)	81.93(4)	87.72(3)	76.85(5)	90.54(1)

Table 2**: F-measure of different algorithms (%)**

Dataset ↓/ Algorithm→	AWE	SERA	SMOTE	OOB	CDDA
Hyperplane (Synthetic dataset)	20.00(5)	33.21(4)	42.61(3)	59.84(2)	85.04(1)
SEA$_s$ (Synthetic dataset)	7.95(5)	27.30(4)	41.37(3)	42.75(2)	48.04(1)
Electricity (Real Dataset)	26.18(4)	18.12(5)	28.42(2)	27.24(3)	32.4(1)

Table 3: G-means of different algorithms (%)

Dataset ↓/ Algorithm→	AWE	SERA	SMOTE	OOB	CDDA
Hyperplane (Synthetic dataset)	31.57(5)	65.61(4)	75.57(3)	89.83(2)	92.4(1)
SEA$_s$ (Synthetic dataset)	16.81(5)	49.54(4)	79.22(3)	80.47(2)	81.42(1)
Electricity (Real Dataset)	48.74(4)	34.67(5)	62.69(2)	62.87(1)	61.67(3)

Table 5: Parameters' of performances of the average ranks of different algorithms on different data sets.

Metric ↓/ Algorithm→	AWE	SERA	SMOTE	OOB	CDDA
Accuracy	**1.3**	3.3	4	4.7	1.7
F-measure	4.7	□□□	2.7	2.3	**1.00**
G-Mean	4.7	4.3	2.7	**1.7**	**1.7**

From the above tables we can notice several observations. First, we can notice that most models are suffering from drop in the performance when the drift is sudden. Second, The AWE is always maintaining a high level of accuracy compared to the other algorithms and this comes at the expense of other parameters which are used for imbalanced data. This is because that the AWE is lacking the way to handle the majority class. Finally, CDDA outperforms the other tested methods in terms of F-measure and G-means. And this is because CDDA can retrieve its performance quickly when the sudden drift has occurred.

In case of the gradual drift on SEAs, we can notice many observations from Tables 2 to 5, which seem to be similar to those observations on the Hyperplane dataset. First the tested algorithm AWE achieves a high performance in case of accuracy and that is because it has the ability for the good prediction on the majority class. However, its performance is dropped in case of F-measure and G-mean. Second in term of F-measure, G-means and AUC, OOB has good performance but at the cost of accuracy. And finally, CDDA performs superiorly when compared to the other comparative methods so CDDA maintains a good balance in terms of accuracy, G-means and F-measure.

For the real dataset electricity, Tables 2 to 4 provide the average of evaluation metrics with all tested methods. The dataset has been converted to an imbalanced data. A few points can be noticed from the above tables: First, SERA has got the best rank in case of accuracy, followed by CDDA. However, on the other hand it can be understood that SERA cannot recognize the minority observations, which yields to an inferior performance on the other metrics as F-measure and G-mean which got the worst rank. Second OOB obtains a stable increase in the performance and it outperforms the contemporary models in terms of G-means; but this comes at the cost of accuracy which drops to rank 5. Similar story is noticed on the SMOTE algorithm which performs quite well in case of accuracy but it achieves inferior performance on the other metrics. Finally, it can be identified that the CDDA achieves the best performance and it outperforms the other contemporary models in F-measure. Meanwhile, it has got the second rank in terms of accuracy.

VII. SUMMARY

In this paper we have addressed the problem of imbalanced data and detecting the concept drift over the data stream. While each issue individually has been well studied by many researchers, the joint issue of concept drift and class imbalance has been receiving an increased concern but it still remains unexplored. In this paper we have presented CDDA model to handle the combined issue of concept drifts and class imbalance and reacting to different types of drift. As the imbalanced data have common challenges with the standard interpretation, we explored different metrics to analyze the performance of the algorithm. The study presents an empirical investigation on real dataset as well as synthetic datasets. The performance of the CDDA is analyzed. Then we compared the results obtained with the other modern methods in different metrics which are used for imbalanced data like F-measure and G-mean. The experiment proved that for the metric of F-measures and G-mean of CDDA outperform the other comparative models.

REFERENCES

[1]. M. Sayed-Mouchaweh, Learning from data streams in dynamic environments. Springer International Publishing., 2016.

[2]. B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, “Ensemble learning for data stream analysis: A survey,” Inf. Fusion, vol. 37, pp. 132–156, 2017.

[3]. A. L. D. Rossi, B. F. De Souza, C. Soares, and A. C. P. D. L. F. De Carvalho, “A guidance of data stream characterization for meta-learning,” Intell. Data Anal., vol. 21, no. 4, pp. 1015–1035, 2017.

[4]. H. He and E. Garcia, “Learning from imbalanced data,” Data Eng. IEEE Trans., vol. 21, no. 9, pp. 1263–1284, 2009.

[5]. Abdualrhman, Mohammed Ahmed Ali, and M. C. Padma, “Deterministic Concept Drift Detection in Ensemble Classifier based Data stream classification Process,” International Journal of Grid and High Performance Computing (IJGHPC) vol. 11, no. 1, 2019.

[6]. R. Polikar, L. Udpa, S. Member, S. S. Udpa, and V. Honavar, “Learn ++ : An Incremental Learning Algorithm,” Comput. J., vol. 58, no. 3, pp. 457–471, 2015.

[7]. X. Hong, S. Chen, and C. J. Harris, “A kernel-based two-class classifier for imbalanced data sets,” IEEE Trans. Neural Networks, vol. 18, no. 1, pp. 28–41, 2007.

[8]. K. W. P. Chawla, N.V., Bowyer,K.W., Hall, L.O., “SMOTE: Synthetic Minority Over-Sampling Technique”. Journal of Artificial Intelligence Research,” vol. 16, pp. 321–357, 2002.

[9]. J. Gama, P. Medas, G. Castillo, and P. Rodrigues, “Learning with Drift Detection,” pp. 286–295, 2004.

[10]. A. Bifet et al., “Early Drift Detection Method,” 4th ECML PKDD Int. Work. Knowl. Discov. from Data Streams, vol. 6, pp. 77–86, 2006.

[11]. A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, “MOA Massive Online Analysis,” J. Mach. Learn. Res., vol. 11, pp. 1601–1604, 2011.

[12]. H. Chen, S., & He, “sera: selectively recursive approach towards nonstationary imbalanced stream data mining,” in Neural Networks, 2009. IJCNN 2009. International Joint Conference, 2008, no. 201, pp. 1141–1141.

[13]. H. Wang, W. Fan, P. Yu, and J. Han, “Mining concept-drifting data streams using ensemble classifiers,” Ninth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 226–235, 2003.

[14]. K. W. P. Chawla, N.V., Bowyer,K.W., Hall, L.O., “SMOTE: Synthetic Minority Over-Sampling Technique”, Journal of Artificial Intelligence Research,” vol. 16, pp. 321–357, 2002.

[15]. S. Wang, L. L. Minku, and X. Yao, “Resampling-based ensemble methods for online class imbalance learning”, IEEE Trans. Knowl. Data Eng., vol. 27, no. 5, pp. 1356–1368, 2015.