Active learning from online data streams with imbalance data: An overview

Anupama N, Sudarson Jena

Abstract— **The traditional classification approaches proposed for knowledge discovery are suitable for learning from the stereo typed data sources which are of uniform in classes. The data available in the real world is not only of imbalance in nature but also the data received is in the form continuous chunks for knowledge discovery. The current researchers are challenged with the issues of data streams which are imbalance in nature. The problem of class imbalance learning and data streams are addressed independently by many researchers. In this paper, we presented an elaborated survey of various recent proposals in this field of research.**

Index Terms— **Class imbalance, Concept drift, Data streams, Classification, Decision trees.**

I. INTRODUCTION

Learning how to classify objects to one of a pre-specified set of categories or classes is a characteristic of intelligence that has been of keen interest to researchers in psychology and computer science. Identifying the common —core characteristics of a set of objects that are representative of their class is of enormous use in focusing the attention of a person or computer program. For example, to determine whether an animal is a zebra, people know to look for stripes rather than examine its tail or ears. Thus, stripes figure strongly in our *concept* (generalization) of zebras. Of course stripes alone are not sufficient to form a class description for zebras as tigers have them also, but they are certainly one of the important characteristics. The ability to perform classification and to be able to *learn* to classify gives people and computer programs the power to make decisions. The efficacy of these decisions is affected by performance on the classification task.

In machine learning, the classification task described above is commonly referred to as *supervised learning*. In supervised learning there is a specified set of classes, and example objects are labeled with the appropriate class (using the example above, the program is told what a zebra is and what is not). The goal is to generalize (form class descriptions) from the training objects that will enable novel objects to be identified as belonging to one of the classes. In contrast to supervise learning is *unsupervised learning*. In this case the program is not told which objects are zebras. Often the goal in unsupervised learning is to decide which objects should be grouped together—in other words, the learner forms the classes itself. Of course, the success of classification learning is heavily dependent on the quality of the data provided for training—a learner has only the input to learn from. If the data is inadequate or irrelevant then the concept descriptions will reflect this and misclassification will result when they are applied to new data. The popular approach of classification algorithms are C4.5, CART and REP.

II. RECENT ADVANCES IN DATA STREAM MIMING

The recent contributions in the field of data stream area as follows,Vladimir Nikulin *et al.* [1] have ranked the given field of customers in accordance to their loyalty or intension to repurchase in the near future. Xiaoxuan Zhang *et al.* [2] have propose an asymmetric active querying strategy that assigns different probabilities for query to examples predicted as positive and negative. Roberto L. Shinmoto Torres *et al.* [3] have proposed a class-wise dynamically weighted CRF (dWCRF) where weights are automatically determined during training by maximizing the expected overall F-score. P.K.srimani *et al.* [4] have aimed at mining data streams with concept drift in Massive Online Analysis Frame work by using Naive Bayes algorithm using classification technique.

Ireneusz Czarnowski *et al.* [5] have proposed and validate a new approach to mining data streams with concept-drift using the ensemble classifier constructed from the one-class base classifiers. In the proposed approach each chunk consists of prototypes and can be updated using instance selection technique when a new data have arrived. Dariusz Jankowski *et al.* [6] have proposed algorithm, named Concept-adapting Evolutionary Algorithm for Decision Tree does not require any knowledge of the environment such as numbers and rates of drifts. The novelty of the approach is combining tree learner and evolutionary algorithm, where the decision tree is learned incrementally and all information is stored in an internal structure of the trees' population. Georg Krempl *et al.* [7] have presented a discussion on eight open challenges for data stream mining. Their goal is to identify gaps between current research and meaningful applications, highlight open problems, and define new application-relevant research directions for data stream mining.

T. Ryan Hoens *et al.* [8] have present an overview of each of the challenging areas in data stream mining, followed by a comprehensive review of recent research for developing of a general framework. Yu Sun *et al.* [9] have proposed a class-based ensemble approach, namely Class-Based ensemble for Class Evolution (CBCE). By maintaining a base learner for each class and dynamically updating the base learners with new data, CBCE can rapidly adjust to class evolution. A novel under-sampling method for the base learners is also proposed to handle the dynamic class-imbalance problem caused by the gradual evolution of classes. Ge Song *et al.* [10] have proposed a new ensemble framework, clustering forest, for learning from the textual imbalanced stream with concept drift (CFIM).The CFIM is based on ensemble learning by integrating a set of clustering trees (CTs). An adaptive selection method, which flexibly chooses the useful CTs by the property of the stream, is presented in CFIM.

Anupama N,, Research Scholar, GITAM University, Hyderabad, India.
Sudarson Jena, Sambalpur University Institute of Information Technology, Sambalpur, Odisha, India

Mansurul A. Bhuiyan *et al.* [11] have proposed a frequent sub graph mining algorithm called Frequent sub graph mining for huge demand (FSM-H) which uses an iterative Map Reduce based framework. FSM-H is complete as it returns all the frequent sub graphs for a given user-defined support, and it is efficient as it applies all the optimizations that the latest FSM algorithms adopt. Xindong Wu *et al.* [12] have presented a HACE theorem that characterizes the features of the Big Data revolution model, from the data mining perspective. Hao Zhang *et al.* [13] have presented a review of a wide range of in-memory data management and processing proposals and systems, including both data storage systems and data processing frameworks

R. J. Lyon *et al.* [14] have presented one possible data processing scenario for the Square Kilometre Array (SKA), for the purposes of an all-sky pulsar survey. In particular they treated the selection of promising signals from the SKA processing pipeline as a data stream classification problem. In the proposed approach the feasibility of classifying signals that arrive via an unlabelled and heavily class imbalanced data stream, using currently available algorithms and frameworks is determined. Meenakshi Anurag Thalor *et al.* [15] have developed an ensemble based classification algorithm for non-stationary data stream (ENSDS) with focus on two-class problems of advertisement recommendation system, in which customer's behaviour may change depending on the season of the year, on the inflation and on new products made available.

P.K. Srimani *et al.* [16] have presented a data stream mining approach with concept drift in Massive Online Analysis Frame work by using Naive Bayes algorithm for classification. Gregory Ditzler *et al.* [17] havedescribed two ensemble-based approaches for learning concept drift from imbalanced data. First approach is a logical combination of Learn++.NSE algorithm for concept drift, with the well-established SMOTE for learning from imbalanced data. Second approach makes two major modifications to Learn++.NSE-SMOTE integration by replacing SMOTE with a sub-ensemble that makes strategic use of minority class data; and replacing Learn++.NSE and its class-independent error weighting mechanism with a penalty constraint that forces the algorithm to balance accuracy on all classes. The primary novelty of this approach is in determining the voting weights for combining ensemble members, based on each classifier's time and imbalance-adjusted accuracy on current and past environments. Bartosz Krawczyk [18] have discussed open issues and challenges such as classification, regression, clustering, data streams, big data analytics and applications, e.g., in social media and computer vision in the field of imbalance learning.

III. Real time applicable areas of data streams

3.1 *Financial Fraud Detection:* Dataset of financial transactions comes under the category of data streams and the basic type of transactions available are of two categories one legitimate and other fraudulent. The fraudulent transactions are of minority in class.

3.2 *Industrial System Monitoring:* In the industrial production, the quality of the product is to e monitored. The industrial monitored system encounters the rare cases of fault detection which is a minority class imbalance problem.

3.3 *Network Intrusion Detection:* In the case of network intrusion detection, the data arrive is of the data stream category. The malicious connections are of minority in count when compared to normal connections requested by the clients.

3.4 *Medical Fraud Detection:* In medical fraud detection datasets also two categories exists. One is the genuine transactions, which is majority in class and other is fraudulent transactions, which are minority in class. The detection of minority fraudulent transactions is of very importance.

3.5 *Real Time Video Surveillance:* The real time video surveillance data stream is also in the class imbalance nature. The classes in the video surveillance data stream are suspicious and non suspicious class. The suspicious class is a majority class and non suspicious class is a minority class.

3.6 *Oil Spillage:* The oil spillage images obtained from satellite radar are of imbalance in nature because the accident of oil spillage occurs rarely and this minority class is of very important due to its financial or ecological impact.

3.7 *Astronomy:* The data gathered from astronomical field also consists of rare objects minority class. This minority class of rarely discovered objects leads to may scientific discoveries.

3.8 *Product Recommendation:* The problem of product recommendation is a binary class imbalance problem. The opinions of the users about a product will mostly tends to onside either buy or not to buy.

3.9 *Spam Image Detection:* Duplicate spam image identification dataset is also consists of rare detection class as the minority class.

3.10 *Text Classification:* In text classification also many type of imbalance sub datasets can exists such as class size, text number etc. for efficient classification.

3.11 *Health Care:* The field of health care analysis is one of the crucial fields due to life threatening conditions. The disease class in the dataset is a minority class and needs to be interrupted in an efficient way for life saving of the patience.

IV. Evaluation Criteria's for Class Imbalance Learning

To assess the classification results we count the number of true positive (TP), true negative (TN), false positive (FP) (actually negative, but classified as positive) and false negative (FN) (actually positive, but classified as negative) examples. It is now well known that error rate is not an appropriate evaluation criterion when there is class imbalance or unequal costs. In this paper, we use AUC, Precision, F-measure, TP Rate and TN Rate as performance evaluation measures.

Let us define a few well known and widely used measures. Apart from these simple metrics, it is possible to encounter several more complex evaluation measures that have been used in different practical domains. One of the most popular techniques for the evaluation of classifiers in imbalanced problems is the Receiver Operating Characteristic (ROC) curve, which is a tool for visualizing, organizing and selecting classifiers based on their tradeoffs between benefits (true positives) and costs (false positives).

The most commonly used empirical measure, accuracy does not distinguish between the number of correct labels of different classes, which in the framework of imbalanced problems may lead to erroneous conclusions. For example a classifier that obtains an accuracy of 90% in a dataset with a degree of imbalance 9:1, might not be accurate if it does not cover correctly any minority class instance.

$$ACC = \frac{TP + TN}{TP + FN + FP + FN}$$

Because of this, instead of using accuracy, more correct metrics are considered. A quantitative representation of a ROC curve is the area under it, which is known as AUC. When only one run is available from a classifier, the AUC can be computed as the arithmetic mean (macro-average) of TP rate and TN rate:
The Area under Curve (AUC) measure is computed by,

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \quad \text{Or}$$

$$AUC = \frac{TP_{RATE} + TN_{RATE}}{2}$$

On the other hand, in several problems we are especially interested in obtaining high performance on only one class. For example, in the diagnosis of a rare disease, one of the most important things is to know how reliable a positive diagnosis is. For such problems, the precision (or purity) metric is often adopted, which can be defined as the percentage of examples that are correctly labeled as positive:

The Precision measure is computed by,

$$\Pr ecision = \frac{TP}{(TP) + (FP)}$$

The F-measure value is computed by,

$$F - measure = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call}$$

To deal with class imbalance, sensitivity (or recall) and specificity have usually been adopted to monitor the classification performance on each class separately. Note that sensitivity (also called true positive rate, TP rate) is the percentage of positive examples that are correctly classified, while specificity (also referred to as true negative rate, TN rate) is defined as the proportion of negative examples that are correctly classified:

The True Positive Rate measure is computed by,

$$TruePositiveRate = \frac{TP}{(TP) + (FN)}$$

The True Negative Rate measure is computed by,

$$TrueNegativeRate = \frac{TN}{(TN) + (FP)}$$

V. Challenges and Future trends of learning in data streams

5.1 Challenges of data stream algorithms:

In general learning from skewed data streams is challenging due to following issues.
1. Evaluation Metric:
The traditional evaluations metric used for classification and clustering are not suitable for learning from skewed data stream. The evaluation metrics such as AUC, F-measure, Sensitivity, specificity, recall etc are required for clear distinction in the class imbalance learning.
2. Lack of minority class data for training:
In the scenario of imbalance data learning, the instances of minority class are very limited for building the proper model for prediction of use instances.
3. Treatment of minority class data as noise:
The ratios of instances in the minority class are very less when compared to the ratio of instances i the majority class. Due to the above reason the minority instances are considered as noise for removal and lot of valuable data is lost.
4. Massive and arrive at varying speeds:
In data streams the instances used for model training and testing are massive in volume. The massive volumes of instances are to be handled in an effective way. The speed of data arrival for data streams also vary at different time periods, these all issues are to e handled in an effective way.
5. Data streams also undergo concept evolution considerably over time:
The imbalance data streams comes with another scenario know as concept drift. In concept drift, the minority class changes into majority class and majority class changes into minority class. In this scenario, the process of knowledge discovery becomes a tedious task.
6. Streamed data management using pre-processing:
The stream datasets management is a difficult task, due to its arrival in the form of data chunks. Apply proper pre-processing technique can lead to make the data better in all terms for proper knowledge discovery.
7. Unavailability of Information on timely basis
The data streams are not available at a time for proper knowledge discovery. The data arrives in the form of data chunks on the timely basis and the problem becomes worse when the data arrives as the class imbalance form with the concept drift.
As per the above points we can see that classification of the skewed data streams is and multi-fold problem. All the above issues need to be addressed so as to design an appropriate learning algorithm for skewed data streams.

5.2 Future Trends of data stream algorithms:

Following open issues are of high importance when designing new algorithms for learning from imbalanced data streams:
1. Making models simpler, more reactive, and more specialized.
2. Minimizing parameter dependence.
3. Combining offline and online models.
4. Solving the right problem using domain knowledge.
5. Developing methods for ensuring privacy.
6. Developing models that handle incomplete, delayed and/or costly feedback.
7. Taking advantage of relations between streaming entities.

8. Developing event detection methods and predictive models for censored data;
9. Developing a systematic methodology for streamed pre-processing.
10. Creating simpler models through multi-objective optimization criteria, which consider not only accuracy, but also computational resources, diagnostics, reactivity, interpretability.
11. Developing online monitoring systems, ensuring reliability of any updates, and balancing the distribution of resources.
12. Unavailability of class labels – Absence of class labels lead to develop labelling strategies – Clustering techniques.
13. Using characteristics and structure of minority class is a promising direction for static imbalanced learning.

VI. Concluding Remarks And Future Work

In this paper, we presented an elaborated survey of various recent proposals in the field of imbalance learning and data streams. The practical applicability of the real world data for efficient knowledge discovery is the need of the hour. Regardless of the recent proposals done by many researchers in this field, still there is lot of scope for innovative and novel proposals. Hence, we propose the future work to be directed towards developing algorithms capable of efficient knowledge discovery from imbalance data streams.

References

[1] Vladimir Nikulin," PREDICTION OF THE SHOPPERS LOYALTY WITH AGGREGATED DATA STREAMS", JAISCR, 2016, Vol. 6, No. 2, pp. 6 9-79.

[2] Xiaoxuan Zhang, Tianbao Yang, Padmini Srinivasan," Online Asymmetric Active Learning with Imbalanced Data", KDD '16, August 13-17, 2016, San Francisco, CA, USA, 2016 ACM. ISBN 978-1-4503-4232-2/16/08.

[3] Roberto L. Shinmoto Torres, Damith C. Ranasinghe, Qinfeng Shi, and Anton van den Hengel," Learning from Imbalanced Multiclass Sequential Data Streams Using Dynamically Weighted Conditional Random Fields",

[4] P. K. SRIMANI, MALINI M PATIL," MINING DATA STREAMS WITH CONCEPT DRIFT IN MASSIVE ONLINE ANALYSIS FRAME WORK", WSEAS TRANSACTIONS on COMPUTERS, Volume 15, 2016.

[5] IreneuszCzarnowski, PiotrJe¸drzejowicz," Ensemble classifier for mining data streams", 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014. doi: 10.1016/j.procs.2014.08.120.

[6] Dariusz Jankowski and KonradJackowski and BogusławCyganek," Learning Decision Trees from Data Streams with Concept Drift", ICCS 2016. The International Conference on Computational Science, Volume 80, 2016, Pages 1682–1691.

[7] Georg Krempl, Indre ˇZliobaite, DariuszBrzezi ´nski, Eyke H¨ ullermeier, Mark Last, Vincent Lemaire, TinoNoack, Ammar Shaker, Sonja Sievi, Myra Spiliopoulou, Jerzy Stefanowski," Open Challenges for Data Stream Mining Research" SIGKDD Explorations Volume 16, Issue 1.

[8] T. Ryan Hoens, RobiPolikar, Nitesh V. Chawla," Learning from streaming data with concept drift and imbalance: an overview",ProgArtifIntell (2012) 1:89–101, DOI 10.1007/s13748-011-0008-0.

[9] Yu Sun, Ke Tang, Leandro L. Minku, Shuo Wang, Xin Yao, "Online Ensemble Learning of Data Streams with Gradually Evolved Classes", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

[10] Ge Song and Yunming Ye," A Dynamic Ensemble Framework for Mining Textual Streams with Class Imbalance",Hindawi Publishing Corporation Scientific World Journal, Volume 2014, Article ID 497354, 11 pages, http://dx.doi.org/10.1155/2014/497354.

[11] Mansurul A. Bhuiyan and Mohammad Al Hasan "An Iterative MapReduce Based Frequent Subgraph Mining Algorithm",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 3, MARCH 2015.

[12] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE,"Data Mining with Big Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014.

[13] Hao Zhang, Gang Chen, Member, IEEE, Beng Chin Ooi, Fellow, IEEE, Kian-Lee Tan, Member, IEEE, and Meihui Zhang, Member, IEEE,"In-Memory Big Data Management and Processing: A Survey",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 7, JULY 2015

[14] R. J. Lyon, J. M. Brooke, J. D. Knowles, B. W. Stappers," A Study on Classification in Imbalanced and Partially-Labelled Data Streams",

[15] MeenakshiAnuragThalor, ShrishailapaPatil," Incremental Learning on Non-stationary Data Stream Using Ensemble Approach", International Journal of Electrical and Computer Engineering (IJECE).

[16] P. K. SRIMANI, MRS. MALINI M PATIL," MINING DATA STREAMS WITH CONCEPT DRIFT IN MASSIVE ONLINE ANALYSIS FRAME WORK", WSEAS TRANSACTIONS on COMPUTERS, Volume 15, 2016.

[17] Gregory Ditzler and RobiPolikar," Incremental Learning of Concept Drift from Streaming Imbalanced Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Digital Object Indentifier 10.1109/TKDE.2012.136

[18] BartoszKrawczyk," Learning from imbalanced data: open challenges and future directions",ProgArtifIntell, DOI 10.1007/s13748-016-0094-0.

[19] João Gama," A survey on learning from data streams: current and future trends",ProgArtifIntell (2012) 1:45–55,DOI 10.1007/s13748-011-0002-6.

Anupama N currently doing research in GITAM(deemed to be university) Hyderabad. She received B.E in Electronics and Instrumentation from Andhra University in 2005 and M.Tech in Computer Science and engineering from JNTU Hyderabad in 2008. She published seven papers in various national, international journals and conferences. Her research interests include , Data Mining and Wireless Sensor Networks.

Sudarson Jena is currently working as Associate Professor in the Department of Computer Science Engineering, Sambalpur University Institute of Information Technology (SUIIT),Sambalpur , Odisha. He received M. Tech degree in Computer Science and Engineering from JNTU- Hyderabad and Ph.D degree in Computer Science from Sambalpur University, in 2008. Dr. Jena has so far published more than 70 Technical papers in referred Journals and Conference proceedings. Dr. Jena was the Joint-Editor of the Journal, The Chanakya during 2005-2007 and Mentor of Interscience Research Network (IRNet) since 2012.He serves on the editorial board of Asian Journal of Engineering and Technology, International Journal of Computer Science Engineering, American Journal of Intelligent Systems, Journal of Grid and Distributed Computing and ten other journals. Dr Jena is an invited speaker and acted as a Chairperson and Co-Chairs for various International, National Conferences/workshops and other Research groups. He is a Senior Member of IEEE and Life Member of Computer Society of India (CSI) and ISTE and senior member of International Association of Computer Science and Information Technology (IACSIT). His research interests include Parallel and Distributed System, Data Mining, Reliability and Performance Evaluation of Interconnection Networks, Grid Computing, Cluster Computing and Soft Computing.