ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 5, May 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

A Survey on Approaches to Efficient Classification of Data Streams using Concept Drift

Aditee Jadhav[1]
Dept. of Computer Engineering,
VIIT,
Pune – India

Leena Deshpande[2]
Dept. of Compute Engineering,
VIIT,
Pune – India

***Abstract:** Recently advancement in hardware and software has enabled processing of large amount of data efficiently. Many applications generate big data rapidly in high fluctuating rates. The sequence of instances arriving at the rate which does not allow it to be stored in memory is referred as data stream. This stream data is generated continuously and at such rate that it is not possible to store them on disk and then process on them. This necessitated techniques for mining data streams on fly using less possible storage space while maintaining low running time. Traditional data mining techniques are not sufficient for data streams, thus introducing stream data mining as a new field. Stream data mining has introduced many challenges such as size of data streams, possible concept drifts. Classification of stream data is a crucial task which needs to address these challenges. Traditional approaches of classification alone are not efficient in the cases when underlying distribution generating data streams changes over time. This induces erroneous results of classification if changes in data streams are not handled with special techniques. This paper focuses on the study of techniques of classification using concept drift and their merits and limitations.*

***Keywords:** Data stream mining; concept drift; ensemble.*

I. Introduction

Due to increasing use of computer applications, large amount of data is generated for storage and processing. From credit card transactions to stock indexes, from blog posts to network traffic, modern applications record massive datasets which may contain hidden patterns and knowledge. The extraction of this knowledge manually cannot be done due to volume of gathered data. This leads to development of data mining methods that automatically discover interesting patterns hidden in vast datasets. Association mining, clustering and classification are traditional data mining tasks, which have been perfected over last two decades. But these techniques are usually applied to complete, static datasets. Recently applications generate massive data in the form of data streams, processing of which faces many problems.

A data stream is unbounded sequence of instances, which arrive at such high rate that does not allow to store them permanently and process them again. Applications where data needs to be processed in the form of streams are sensor networks, stock exchange transactions, banking, and many more. The presence of data streams in data mining technology leads to emergence of novel approaches of data mining called data stream mining.

Data stream learning faces three challenges: size, speed and variability. The first two challenges force data stream algorithms to process data instances in limited amount of time and memory. Variability means changing patterns in dynamic environment. This refers to concept drift meaning changes in the distribution and definition of underlying concept. Such unpredictable changes in incoming data instances hamper the accuracy of classification model trained from past instances. Therefore, for dealing with concept drifts, data mining methods implement adaptation or drift detection techniques to cope with changing environment.

In data stream mining, classification is the most studied task. Classification process aims at generalizing the known facts presented in data instances and applying this knowledge to new data. Ensemble classifier is most common technique used for dealing with data streams. Ensemble is set of single classifiers, whose predictions are aggregated to produce final output. The modular nature of ensemble provides natural way to adapt changes by modifying their structure. This paper focuses on study of classification techniques for data streams as well as for dealing with concept drift.

II. Data Stream Classification

Classification is the task of assigning data instance to one of predefined classes. The input for classification is data instance characterized by a tuple { x,y } , where x is set of attributes describing data instance and y is class label of data instance.

Definition 2.1: Classification is the task of learning function that maps each attribute set x to one of predefined class labels y.

Classification technique is used in context of predictive modelling. The examples of classification include spam detection based on email text, outlier detection based on previously studied sample data. To choose best possible model for classification task, evaluation criteria is needed. The main factor used while choosing classifier is predictive performance analyzed by simple error-rate. Error rate is the fraction of misclassified instances, while accuracy is the fraction of correctly classified instances. The classifier model is built using testing data, while accuracy is measured based on testing data. Traditionally, classification task is evaluated using static datasets.

Definition 2.2: Data stream is infinite, ordered sequence on instances x^t (t=1,2,...T) that arrive at such rate that does not allow them to be stored and then processed.

There are several constraints on classification due to speed and size of data streams:

1. All the data from data streams cannot be stored in memory. Summaries of data are computed and stored and rest of information is discarded
2. Due to speed of arriving data instances, data stream need to be processed only once.
3. The distribution that generates data streams may change over time, making past data irrelevant or harmful for processing of current data.

There are two ways of processing data streams: incrementally (online processing) or in portions (block processing). The first approach processes instances one by one as they appear in consecutive moments of time. In this type, label y^t of instance x^t is available before arrival of instance x^{t+1}. This approach is also called instance incremental processing. In the later approach, instances are processed in larger sets called blocks.

III. Concept Drift

One of the most important properties of data streams is that the underlying concept may change over time. Thus approach to data stream classification needs to be capable of detecting and adapting concept changes. The term concept drift refers to the fact that the concept about which data is being collected may change over time after certain stability period. Concept drift is detected when probability distribution generating data streams change. Experiments on classification of data streams using concept drifts have shown that changes in underlying concept deteriorate performance of classifier model. This necessitates study of concept drifts and techniques to handle them. There are two broad categories of concept drift depending upon how drift occurs: sudden drift and gradual drift.

Sudden drift is said to occur when source distribution p^t at time t is suddenly replaced by other distribution at time t+1 . This abrupt drift degrades performance of classifier as the model has been trained with instances of other distribution. Gradual drift is associated by slower rates of changes. This type of drift is observed when instances from different distributions p^t and

$p^{t+\Delta}$; are mixed. As time goes on, probability of observing instances from new type of distribution $p^{t+\Delta}$; increases. Sudden drift is detected once performance of classifier deteriorates all of sudden. But gradual changes are difficult to be detected and analyzed.

When no information is available about distribution of data, the problem of concept drift can be solved by monitoring performance of classifier by using classification accuracy as performance measure. Decay of classifier accuracy below predefined threshold value is used for signalling concept drift. Classifier ensemble is most popular way of dealing with such problems. Ensemble adapt to changes by modifying its base components or aggregation method..

Ensemble is set of single classifiers whose decisions are aggregated to produce final decision for given data instances. The combined decision of many classifiers is more accurate than that given by single classifier. The components of ensemble differ from each other depending on instances they have been trained on, the attributes they used for building model or classification technique they use. Commonly aggregation of decisions is done using majority voting. Furthermore, for dealing with concept drifts, ensemble incorporates forgetting mechanism. This is achieved by adding new classifier after each block of instances, by changing weights of ensemble components or by replacing weakest component with classifier trained on most recent block of data. There are two types of ensemble classifiers for data stream classifications: online ensemble, where model learns incrementally after processing each instance and block ensemble that processes when block of instances is available. Online ensemble is best suited for dealing with sudden drift, as it is updated after each instance of data is available, whereas block ensemble work best on gradual drift.

IV. Related Work

Existing techniques of data stream classification handle concept drift and efficiency aspects of classification process. Each of these techniques follow incremental or batch approach for classification problem. There are two variations observed in these approaches. The first approach is classification model for dealing with efficient classification of data streams using concept drifts. Several classification algorithms have been developed to deal with concept drift in data streams. But they target at one of the concept drifts: sudden, gradual or recurring. Brzezinski et. al. [2] proposed Accuracy Updated Ensemble(AUE2) that works equally well on different types of drifts. The system maintains weighted pool of base classifiers and predicts class of incoming data instance by aggregating predictions of base classifiers by weighted voting rule. After classification, weights of base classifiers are updated based on accuracy of classification. For each new data chunk, new base learner is constructed and least accurate classifier is discarded. Because new classifier is built for each new chunk, gradual drift can also be detected well. Ensemble models build classifiers that require intensive labelling of classes. But a small number of labelled instances are available for training classifiers. A large number of unclassified instances can be used to build clusters from data streams. But clusters are insufficient for classification, as they carry no genuine class label information. Zhang et. al. [12] proposed ensemble of both classifiers and clusters to address this issue. The system first builds classifiers using training data. Then for each incoming data chunk, clusters are formed and class label information is propagated from classifiers to clusters. By mapping similarity between class label information and cluster ID, class labels are assigned to data instances. Clusters are split into clusters constituting instances with same class label. The unlabeled instances form separate cluster which represent concept drift.

Analysis of spatio-temporal data is done generally using sentiments in geo-tagged tweets of the users. This analysis can be used for detection of spatial trajectories of moving objects. The identified trajectories help analyst to find interesting patterns as well as any drift occurred in patterns. Senaratne et. al. [1] provided a framework that utilizes kernel density estimation for detecting hotpot clusters of twitter activities. Traditionally online learning algorithms have been used for dealing with concept drifts. But they do not maintain high generalization level on both old and new concepts. To solve this issue, Minku et. al. [5] proposed a system that keeps different ensembles for different diversity levels to attain better accuracy. The learning system is composed of two ensembles: ensemble with low diversity and ensemble with high diversity. When drift is detected, again two ensembles are created depending upon the level of diversity. Use of old ensembles and new ensembles helps to maintain high generalization of old and new concepts and provides high accuracy, as information in old concept is used for learning new concept.

The other approach deals with concept evolution or novelty detection, in addition with concept drift. Concept evolution occurs when new classes are introduced in data streams. Masud et. al. [11] Handled concept evolution problem by proposing adaptive threshold for detecting outliers. A k-NN based classifier is trained for each chunk of data stream. Classified data is clustered using k-means clustering and summaries for clusters are saved. These clusters make hypersphere by union among themselves which makes a decision boundary. Generally data instances falling outside the boundary are detected outlier. But sometimes data instances belonging to existing class may fall outside the boundary, but very close to the surface of hypersphere, therein resulting into high false alarm rate. This system proposed slack surface outside hypersphere. Data instances lying outside slack surface are categorized as outliers. Novelty score is computed for outlier to measure separation of outliers from existing class and the cohesion among outliers. All novelty scores are discretized into intervals and Gini coefficient is computed for intervals. Depending upon values of Gini coefficient, novel class is declared. Novelty detection and concept drift has introduced new challenges in classification of data streams. Memory requirement and storage space are limited for handling these challenges. Hayat et. al. [13] proposed compact model based on clustering algorithms. Previously, for clustering (distance based) techniques, whole cluster was considered to evaluate input data instance. Hayat suggested the use of only neighbourhood of instance. The system builds normal clusters using k-means on training data. These clusters are partitioned into equal sized sub clusters by applying k-means on each cluster. Discrete cosine transform is applied on sub clusters to generate k-neigh coefficients. K-neigh is the number of neighbours to be considered for evaluating input instance. Classification is performed using Euclidean distance and selecting closest cluster to input instance. The instances that are not classified are considered as unknown data. K-means is applied on unknown data producing clusters for each of which, DCT is also applied. These clusters are compared with normal sub clusters. If dissimilarity is more than certain threshold, it is considered novelty, else it is concept drift.

Novel classes, concept drifts and feature evolution have imposed challenges to ensemble approach for mining stream data. Classifiers in ensemble need to be updated incrementally. This involves processing even on irrelevant features. This problem is addressed by Hierarchical stream miner proposed by Khan et.al. [6]. HSMiner is hierarchical decomposition approach to the ensemble classifier. The main multi class ensemble is decomposed into multiple single class ensembles. Each single class ensemble constructs sub learners for each feature of that class. For new instance passed down the hierarchy, votes are collected from base learners to learn class of instance or to detect outlier. These outliers are fed to separate novel class learner trained using pseudo data. Cohesion among outliers is checked for separating noisy mix of instances and labelling novel classes.

V. Future Work

The above contributions open several directions for research studies. Current works in data stream classification focus on single type of concept drift. Combining different types of drifts lead to new line of research. Also more approaches of classification can be developed for improving performance with least training data usage. An interesting future work would be to identify evolution of new concepts. Recent techniques can be further extended for solving novelty detection problem.

VI. Conclusion

Rapid generation of big data has introduced data stream mining techniques to the existing world of data mining. Data streams being generated continuously at high fluctuating rates, are processed on the fly. It is not feasible to store them and then process, since distribution of generated data changes over time. These changes must be updated regularly to reflect current data streams. There are certain changes in data streams which describe patterns of interest. Such changes are detected by finding dissimilarity between old summarized data and newly arrived data.

Ensemble classifier is most commonly used approach for handling concept drifts in data streams. Online ensemble, that learns as soon as data instance is available, helps in detecting sudden drifts. Block ensemble learns block of instances at a time and works well on data streams with gradual drifts. This paper gives details of approaches proposed for dealing with concept drifts.

ACKNOWLEDGEMENT

The authors are grateful to reviewers for their comments and suggestions that helped us to improve the paper.

References

1. Senaratne,Lehle, “Moving on Twitter: Using Episodic Hotspot and Drift Analysis to Detect and Characterize Spatial Trajectories”, ACM, Nov 2014.
2. Brzezinski,Stefanowski, “Reacting to Different Types of Concept Drift: Accuracy Updated Ensemble Algorithm”, IEEE, Vol. 25, No. 1, January 2014.
3. Zliobaite, Bifet, “Active Learning with Drifting Streaming Data”, IEEE, Vol. 25, No.1, January 2014.
4. Masud, Chen, Aggarwal, “Classification and Adaptive Novel Class Detection of Feature Evolving Data Streams”, IEEE, Vol. 25, No. 7,July 2013.
5. Minku, Yao,“DDD: A new ensemble approach for dealing with concept drift”, IEEE, Vol. 24, No. 4, April 2012.
6. Parker, Mustafa, Khan, “Novel class detection and feature via a tiered ensemble approach for stream mining”, IEEE, Vol. 16, No. 8, November 2012.
7. Zhu, Zhang, Lin, “Active learning from stream data using classifier ensemble”, IEEE, Vol. 40, No. 6, December.
8. Gao, Mausud, Han, “Classification and novel class detection in concept-drifting data streams under time constraints”, IEEE, Vol. 23, No. 6, June 2011.
9. Masud, Chen, Khan, “Integrating Novel Class Detection with Classification of Concept Drifting Data Streams”, ECML, Springer/PKDD 2010.
10. Cabanes, Bennani, “Change Detection in Data Streams through Unsupervised Learning”, IEEE World Congress on Computational Intelligence June 2012.
11. Masud, Chen, “Addressing Concept-Evolution in Concept Drifting Data Streams”, IEEE International Conference on data mining, 2010.
12. Zhang, Zhu,“Classifier and cluster ensemble for mining concept drifting Data Streams”, IEEE International Conference on Data Mining, 2010.
13. Hayat, Hashemi, “DCT Based Approach for Detecting Novelty and Concept Drift in data streams”, IEEE Conference on Soft Computing and Pattern Recognition, Dec. 2010.
14. Folino, Pizzuti, “An Adaptive Distributed Ensemble Approach to Mine Concept-drifting Data Streams”, IEEE Conference on Tools for Artificial Intelligence, 2007.

AUTHOR(S) PROFILE

Aditee Jadhav, received the B.E. degree in Information technology from Sant Gadge Baba Amravati University, in 2014. She is currently pursuing the M.E. degree in Computer Engineering from Savitribai Phule Pune University. Her current research interests include data stream mining, concept drift.

Mrs. Leena Deshpande, received M.E. in Computer Engineering in 2006 and is an Assistant Professor in Department of Computer Engineering, Vishwakarma Institute of Information Technology. Her current research interests include information retrieval, data mining, and Big data.