Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

(Special Issue for ITECE 2016)

A Review on Limited Labeled Data in Data Stream

Hiral Desai[1], Jay Gandhi[2]

[1]*Computer Engineering Department, B.H.Gardi College Of Engineering and Technology*
[2] *Computer Engineering Department, B.H.Gardi College Of Engineering and Technology*

Abstract - Assigning labels to unlabeled data in data streams from the few labeled data is a momentous and interesting issue for machine learning and data mining that commonly looks in real world stream classification hitches. Till now there are few research work has been done to labeled the unlabeled data for stream data, but Most existing work on classification of data streams take that all streaming data are labeled and the class labels are promptly available. However, in real-life uses, such as fraud and intrusion detection, this suspicion is not generally valid. It has more costly and time consuming to labeled the data manually. Accuracy of labeling the data stream is between 34% to 95% . But for image dataset of stream data we have maximum accuracy is 78%. In proposed approach with the use of SVM (support vector machine) and co-relation to the image type of stream data we can improve accuracy for labeling.

Keywords -*Data stream mining; Classification; Semi-Supervised Learning; Limited Labeled Data; Clustering.*

I. INTRODUCTION

All the time data comes as streams (continuous data). Pleasing vast volumes of streaming data in the fundamental memory is unreasonable and infeasible. Hence, we have to do online processing of streaming data. In this situation, projecting models should be trained either in incremental approach by constant update or by retraining using current batches of data [1].

Data Stream Mining is the procedure to mine the knowledge structures from unremitting and speedy data records. A data stream is an Continuous sequence of instances/examples that in many applications of data stream mining can be read only once (or a small number of times) using limited computing power and limited memory storage abilities. Examples of such stream data are Traffic in Computer network, telephonic conversations, Bank or ATM's transactions, Online Shopping, differ rent Web searches, Weather prediction. It can be consider as a subfield of KDD (knowledge discovery process) and data mining [2].

There are mainly four challenges: Infinite length, Concept drift, Concept evolution, limited labeled data [3][4]. Here we discussed on limited labeled data. The speed at which data points are labeled tags far behind the speed at which at data points arrive in the stream [7][9]. The comments of labeled examples are every now and again tedious and once in a while difficult to gain in some genuine issues. In this way, a great classifier in streaming environment ought to have the capacity to concede the training until genuine names get to be accessible yet keeping labeling recently arrived examples utilizing the current classifier [3]. Also, it ought to have the capacity to utilize mostly named training data.

Active learning and semi-supervised learning have been proposed as an option way to deal with illuminate limited labeled data which mutually abuse labeled and unlabeled examples for training classifiers to extending arrangement exactness. Classification and clustering based on active learning methods such as Support Vector method, Bayesian rule and neural network. Some Semi-supervised learning models are self-training, mixer models, graph based methods, co-training, multi-view learning for limited labeled data [5].

The rest of this paper is organized as takes after. Segment 2 presents related deals with limited labeled data. Area 3 clarifies our proposed technique. Segment 4 presents analyses and last we close our paper in Section 5.

II. RELATED WORK

Because of developing businesses an extensive size of data is produced as script, pictures or exploratory records and commercial exchange and it is hard to collection this tremendous data and examinations it. Knowledge Detection is required to bode well and utilization of these data. This enormous bulk can't be handled straightforwardly as a result of its many-sided quality, for example, repetition, irregularity and so on. It is required to mine information for examination as indicated by our necessities.

Mohammad et al [6] used fixed sized of ensemble classifier to classify the unlabeled data. Proposed model is assembled as micro-clusters utilizing semi-supervised clustering method and classification is performed with κ-nearest neighbor algorithm. Since it requires a fewer amount of labeled data. Proposed SmSCluster algorithm applied on 1600 records for botnet dataset and 1000 records for synthetic dataset with only 5 % labeled points.
Peipei li et al [7] proposed a semi-supervised classification algorithm REDLLA (Recurring concept Drifts and Limited Labeled data) for data stream. REDLLA calculation name the unlabeled data with a clustering approach in the developing of a decision tree and reuse the unlabeled data consolidated with the marked data for split tests of current tree. Furthermore, it utilizes deviation between clusters to recognize new idea and repeating idea at leaves. It applied on both synthetic and real domain recurring concept drifting data. Time consumption of REDLLA algorithm is more. Still there were some challenges like instructions to decrease space consumption, how to alter the times of recurring concept drifts precisely and how to forecast obscure theories ahead of time.

Re-training semi-supervised approach used for enhancing the quality of classifier and also associated with unlabeled data to improve classification accuracy by Urjita thakar et al [8]. Another semi-supervised classifier has been composed which utilizes unlabeled information tests and K-means clustering calculation has been utilized to pick unlabeled examples and name them. These recently marked examples are utilized to re-prepare the classifier incrementally. Firstly, a beginning NN classifier is prepared utilizing some named tests. In parallel, K-means clustering is connected on unlabeled information of the complete dataset. At that point, their clustering marks are balanced by the classifier. The classifier translates the group for this and classifier appoints names to the bunches. The naming specimens whose marks given by the two systems are indistinguishable are decided to re-prepare the classifier and others are dealt with as pointless. Proposed calculation summed up for complete dataset rather than just a section. Improved accuracy of classification on synthetic dataset around 51% and on standard dataset around 37 %. One drawback is there that it increased the execution time.

For Online data stream classification and learning with limited labels using selective self-training semi-supervised classification, an algorithm proposed by Loo Hui Ru et al[9]. Selective self-training method is applied to incrementally learn from both labeled and unlabeled data and the selection of data to be trained can be done as soon as the classification process is complete. Basically proposed method distributed into three portions: offline pre-training, online classification and learning, and cluster reduction. The ability of the proposed method to learn from limited labels is proven by achieving 95% average accuracy by using only 1% labeled data. They used Cambridge and KDD'99 dataset.

UI Jing et al[10] proposed an algorithm ECM-BDF to solve all four challenges of data stream. Firstly, a data stream is distributed into sequential chunks and a classification model is trained from each labeled data chunk. To address the infinite length and concept-drifting issue, a settled number of such models constitute an ensemble model E and consequent labeled pieces are utilized to redesign E. To manage the presence of novel classes and limited labeled instances issue, the model joins a novel class detection mechanism to distinguish the entry of a novel class without training E with labeled instances of that class. For the moment, unsupervised models are trained from unlabeled instances to provide useful constraints for *E*. An extended ensemble model *Ex* can be acquired with the constraints as feedback information, and then unlabeled instances can be classified more accurately by satisfying the maximum consensus of *Ex*.

Classification of satellite images with limited labeled data using SSEP(Semi-Supervised Ensemble Projection) algorithm proposed by Wen Yang et al[11]. Take troupe of powerless preparing (WT) sets inspected from a Gaussian approximation of various element spaces. As data conveyed by a solitary WT set is very restricted, EP receives the thought of outfit realizing, which additionally plans to take in a troupe of classifiers with precision and assorted qualities. The extraction procedure of SSEP depends all in all picture, which is really a scene/object arrangement issue. Diverse classifiers were utilized to assess the elements. They adopted two typical supervised classifiers: LR and linear SVM.

For classification of remote sensing images, co-training semi supervised approach is proposed by Prem Shankar Singh Aydav et al[12]. Effectiveness of the proposed technique, tests have been performed on two diverse spectral perspectives of hyper spectral remote detecting pictures utilizing support vector machine as supervised classifier and semi-supervised fuzzy c-means as clustering technique. For more accurate result compare to traditional classification techniques. In Co-training technique there are using two classifier to train the labeled data. Just most sure specimens are utilized for training yet these examples may not be instructive. This technique achieved approximate 80 to 85% accuracy.

III. PROPOSED WORK

We proposed a semi-supervised classification algorithm to label the unlabeled data and improving classification accuracy to get efficient classification of data. Classification with clustering technique used for classification of stream data. Firstly, all data divides into two perspective like if we consider the data of searching records we can divide into URL through searching and text searching. Other example of image data we can divide based on color and texture. After dividing the data clustering algorithm used for grouping. Train the classifier based on few labeled data and it applies on

test dataset or unlabeled data for labeling data in iterative manner. On two perspective different classifiers or same classifier can used. Here in proposed algorithm use SVM Classifier with correlation coefficient to improve classification accuracy from 80%.

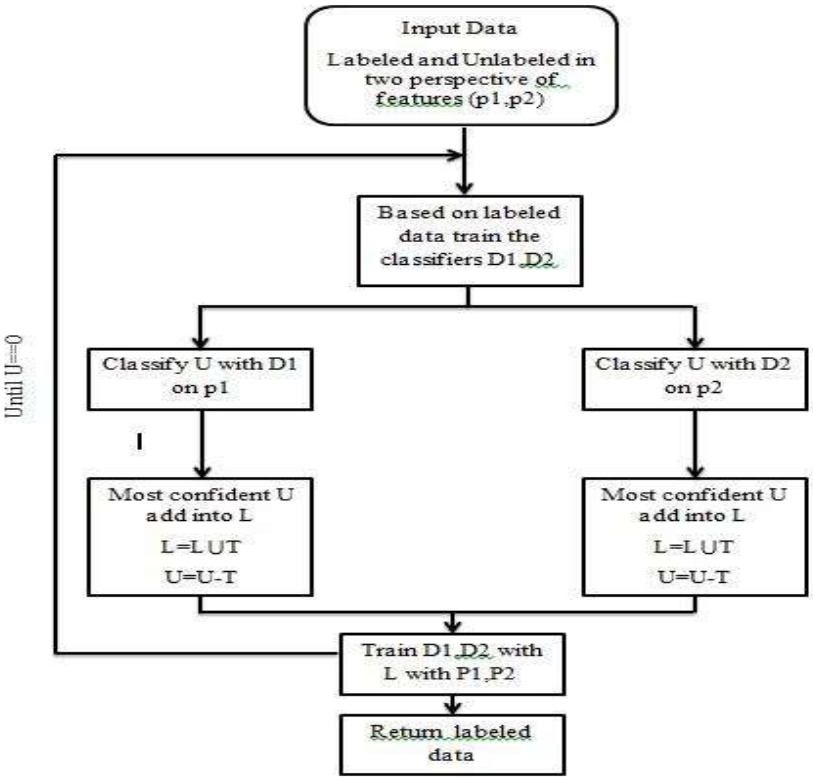

Figure. 1 Flow of proposed algorithm

Input : Label and Unlabeled data pool
With Two Features P1,P2
Output: Labeled data
1.1 Train Classifiers 1,2 with Labeled data
L with Features sets P1,P2
Repeat
1.2 Clustering of the data set with both feature set P1,P2.
1.3 Classify unlabeled data pool U with classifier1 on view p1
1.4 Analysis of data pool U and take most confident labeled data in training set
1.5 Add confidently trained data into training set T
1.6 Classify unlabeled data pool U with classifier2 on viewp2
1.7 Analysis of data pool U and take most confident labeled data in training set
1.8 Add Confidently trained data into training set T
1.9 Train classifier C1,C2 with whole data with Two Features view P1,P2
Until U==0
1.10 Return Labeled data

In proposed approach dataset divided in to two parts or perspectives in step 1.1. From the few labeled data train the classifiers C1 and C2. Then on each perspective apples clustering algorithm for reducing the time complexity of classification in step 1.2. Now in each step from 1.3 to 1.11, iterations of labeled data procedures are there. Classify the unlabeled data using classifier C1 and classifier C2 in step 1.4 and 1.7 respectively. Most confident labeled data take as a training set from the analysis of data after applying classifiers on unlabeled data pool in step 1.5 and 1.8. Respectively in step 1.6 and 1.9 update the classifiers C1 and C2 with newly trained data. Till all unlabeled data not getting labeled iterations will going on and finally we get labeled data set in step 1.12.

IV. EXPERIMENTS

To quantify the exactness of SVM classifier we have made manufactured dataset. We have performed probes a PC with Intel center i3. 1 GB RAM, and windows 8 OS. The datasets have divided into two classes. We obtained following results of accuracy using SVM classifier in every datasets.

Table 1. Result Comparison (Synthetic Dataset)

Sr. No.	Number of instances in dataset	Accuracy
1	11	81.81
2	9	88.88
3	10	90
4	12	91.66
5	60	93.33

V. CONCLUSION AND FUTURE WORK

Labeling the data instances is more time consuming and expansive. There are so many techniques and algorithm are used to labeled the data. But in those accuracy are major concern. Here proposed a semi-supervised classification algorithm for efficient classification of limited labeled stream data. Here focused on image types of data for classification like remote sensed image, satellite images etc. SVM classifier used as a supervised classification algorithm and clustering technique also used as an unsupervised classification algorithm. In future research, With correlation coefficient , SVM classifier improve the classification accuracy and labeled the data to unlabeled data. This algorithm will achieve accuracy more than 80% of image data classification.

VI. ACKNOWLEDGMENTS

The Author might want to thank Assi. Prof. Jay Gandhi of B.H. Gardi college of Engineering and Technology, Rajkot for his insightful discussions.

VII. REFERENCES

[1] [1] Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, "A survey on concept drift adaptation" ACM Computing Surveys 46(4) (2014).

[2] [2] Madjid Khalilian, Norwati Mustapha , "Data Stream Clustering: Challenges And Issues", International Multi-conference for Engineers and Computer Scientists, 2010.

[3] [3] Ms. Priyanka B.Dongre, Dr. Latesh G. Malik, "A Review On Real Time Data Stream Classification And Adapting To Various Concept Drift Scenarios", International Advance Computing Conference (IACC) IEEE-2014.

[4] [4] Mohammad M. Masud, Jing Gaoz, Latifur Khany, Jiawei Hanz, Bhavani Thuraisingham, "Classification And Novel Class Detection In Data Streams With Active Mining"Advance In Knowledge Discovery And Data Mining, Vol.6119, Pp. 311-324, 2010.

[5] [5] Zahra Ahmadi and Hamid Beigy, Semi-supervised Ensemble Learning of Data Streams in the Presence of Concept Drift Springer-Verlag Berlin Heidelberg-2012 , pp. 526–537.

[6] [6] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani Thuraisingham "A Practical Approach To Classify Evolving Data Streams: Training With Limited Amount Of Labeled Data" Eight IEEE International Conference On Data Mining-2008.

[7] [7] Peipei Li, Xindong Wu, Xuegang Hu, "Mining Recurring Concept Drifts With Limited Labeled Streaming Data", 2nd Asian Conference On Machine Learning ,ACM Pp. 241-252 ,2012,

[8] [8] Urjita Thakar, Vandan Tewari, Sameer Rajan, "A Higher Accuracy Classifier Based On Semi- Supervised Learning", International Conference On Computational Intelligence And Communication Networks IEEE Pp. 665-668, 2010.

[9] [9] Loo Hui Ru, Trias Andromeda, M. N. Marsono, "Online Data Stream Learning and Classification with Limited Labels", Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2014)

[10] [10] LIU Jing, XU Guo-Sheng, ZHENG Shi-Hui, XIAO Da, GU Li-Ze, "Data Streams Classification With Ensemble Model Based On Decision-Feedback", The Journal Of China Universities Of Posts And Telecommunications Elsevier Pp. 79-85, 2014.

[11] [11] Wen Yang, Xiaoshuang Yin, Gui-Song Xia, "Learning High-Level Features For Satellite Image Classification With Limited Labeled Samples", IEEE Transactions On Geoscience And Remote Sensing, Vol. 53, NO. 8, Pp. 4472-4482, august 2015.

[12] [12] Prem Shankar Singh Aydav And Sonjharia Minz "Co-Training With Clustering For The Semi-Supervised Classification Of Remote Sensing Images", Proceedings Of The Second International Conference On Computer And Communication Technologies, Springer Pp. 659-667, 2015.