ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

Semi Supervised Learning of Online Data Streams with Max Flow Algorithm

SujataGawade[1], Prof. Vina M. Lomte[2]

P.G. Student, Department of Computer Engineering, RMD Sinhagad Institute of Technology Warje Pune, India[1]

Head of Dept., Department of Computer Engineering, RMD Sinhagad Institute of Technology Warje Pune, India[2]

ABSTRACT: This paper aims to propose the semi-supervised learning system that deals with huge and dynamic data in time and memory efficient manner. In today's world the numbers of applications are available in information system which deals with large amount of data in changing environment. In such environments, this data is available in the form online streams. The memory and time limitations are the major aspects which need to be considered in online learning which are due to volume as well as speed of the data. In online streams, small quantity of information is labelled and huge quantity of unlabelled information is accessible. Consequently, semi regulated learning is the best way to deal with take in the information in order to decrease human efforts and as yet accomplishing better precision also the performance of the system. This paper examined the concept of semi supervised learning. For learning the data stream max flow algorithm is used in this paper. The Max Flow/ Min Cut algorithm is applied for achieving the accuracy of the proposed system. The algorithm shows improved efficiency and enhancement in performance. To further enhance the performance, Cosine Similarity and Feature Selection Techniques are used. The comparative study shows that the proposed system gives better performance in terms of time and memory efficiency than the old learning system. The system is tested using KDD99 dataset for classification of network intrusion attacks. The system shows the higher accuracy and improved performance results. Application: The proposed system gives solutions to various two class problems. The applications such as network intrusion detection system which can be used to classify anomaly and normal network, online audio background noise reduction of streams, traffic bottleneck identification system, image segmentation techniques where image can be segmented in background and foreground, 3D reconstruction and many more can make use of the proposed system.

KEYWORDS: Machine learning, semi supervised learning, Max flow algorithm, data streams, cosine similarity, and feature reduction.

I. INTRODUCTION

Recently, most of the researchers are focus on the examination of large quantity of data, as this method could give the numbers of advantages to the respective organization. One of the critical success factor faced by the most of the enterprises is the opportunity of making the suitable business decision on the basis of hidden knowledge details saved in the organizations. Most of the applications considered that data usually originates uninterruptedly in the form of data stream. Some of the examples of the data streams are analysis of networks, financially data prediction, traffic control, sensor measurement processing, ubiquitous computing, GPS and mobile device tracking, users click log mining, sentiment analysis, and so on. Existing techniques of data mining and machine learning faces the problem because of online data stream, because the existing methods is designed only for static dataset and are not able to do effectively dissecting quickly developing measures of data and taking into deliberation, for example, Limited computational resources as memory and time, and in addition tight needs to make forecasts in sensible time. The phenomenon is concept drift, i.e., changes in dispersion of information which happens in the stream after some time. This could intensely decline execution of the utilized model. Information may come so rapidly in a few applications that labeling all items might be deferred or some of the time even incomprehensible. Researchers focus on the classification concept for overcome the limitations of data stream techniques. It is the ability of computers to learn without being explicitly programmed. In this various algorithms are created so as to facilitate learning. It is based on human learning like

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

learning observations, previous experiences or any other ways. It learns from data as well as makes predictions on data. Due to this, the system can perpetually enhance itself and further offer expanded effectiveness. For the machine learning faced the some difficulties for executing or designing the novel data mining algorithm. Inductive learning and transductive learning are the two types of learning techniques. Inductive learning framework shapes a general rules on set of information. These rule development depends on perceptions. The procedure is otherwise called learning by cases. The framework has no clue about the test information at the time of learning. Moreover, in transductive learning in the researchers manages particular cases consequently it has awareness for test information. The point is to manufacture a characterized for this particular information. Transductive algorithms would label the unlabeled points according to the clusters to which they naturally belong.

Supervised learning provides common approach learning. Supervised learning deals with completely labeled training dataset. The objective involves learning a function from examples of its inputs and outputs so as generalize the unseen data. Every case contains a pair which contains an input value and equivalent desired output value. Another technique is unsupervised learning. In this technique just some example data are offered to the framework as perceptions with no any label. Unsupervised learning utilizes forms that attempt to locate the normal patterns of the information. There is no outside mentor for the framework to find the patterns of the model and it is the individual duty of the researcher to discover the essential activities. In supervised learning the training data is totally marked and in unsupervised adapting, none of the training dataset is labeled.

Semi supervised learning can deals with both labeled and unlabeled information. It enhances the classification accuracy. This technique can enhance the execution of classification with unlabeled information. This strategy is adaptable for all kind of circumstances by distinguishing the connection amongst labeled and unlabeled information. Some case of semi supervised learning calculation incorporates:

- Self-training
- Mixture models
- Graph based techniques
- Co training
- Multi view learning

In graph based learning, the max-flow and min-cut approach is widely used for classifications. The augmenting path approach in1 shows that the classification can be achieved with higher accuracy with minimum amount of labeled instances. In this approach the classification is performed based on Euclidean distance depending on the degree of closeness among the instances assuming the instances which are closer to each other belong to same class. The approach also uses incremental and online learning which means each time a new classifier is found it is update in to learned system. It does not depend only on historical knowledge gained at the time of training the instances for learning instead does incremental learning at the time of testing as well. This approach provides solutions to various two class problems and can be used in various applications such as background noise detection in audio streaming, network anomaly detection, traffic bottleneck identification, image segmentation etc.

II. RELATED WORK

In [1] design and implements the online max-flow algorithm from data streams. First of all implement an online sample network whose sample distance matrix is being improved for every sample adding/retiring, and then more importantly a novel incremental/detrimental max flow algorithm to update the max-flow whenever the sample network changes. An incremental detrimental max-flow algorithm is utilized for learning the online data streams. This

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

system encompasses both labeled and unlabeled data samples. During actual classification, min cut over max flow is computed. This will reduce the number of same kind of sample pairs and classify into unique classes. The performance of system is tested on real word data and proves that this proposed system outperforms other available classification system. However, the system can be improved for time and memory usage by parallelizing some modules and further advanced techniques such as feature selection. The accuracy of the system can be system can be increased by using some advanced similarity measures.

In [2] proposed the Semi Boost semi supervised learning framework. This learning approach improves the performance of any type of classifier. The performance of the system tested on multiple dataset and prove that it improve the accuracy of classification results.

The novel learning algorithm to solve the classification problem in the field of mobile robotics is described in [3]. This system minimizes the amount of interaction with human supervisor, which leads to automatic capacity of learning process. This system has the adaptive nature for every new online incoming data samples. It is flexible in any kind of setting. This system is the combination of online star clustering algorithm and label propagation technique. It can improve the classification accuracy and run time efficiency.

The eT2ELM for human meta cognition system is proposed in [4]. It answers what to learn, how to learn and when to learn. This system also includes type 2 fuzzy neural network model as the cognitive component. In this hidden node is developed with the interval type-2 multivariate Gaussian function. In output node, subset of Chebyshev series is used.

The recent topics are internet learning and Out-of-Sample issues. To take care of these issues, a novel technique named Semi supervised Local Feature Extraction (SLFE) is proposed in [5]. To begin with author extract component in semi-supervised way, at that point characterize a novel complex comparability to develop incremental local tangent space framework. To get multi-complex target work, author use an express mapping function to overcome the Out-of-Sample issue.

A semi-supervised learning technique utilizing an incremental neural system is proposed in [6]. Following the concept drift float is kept up by utilizing incremental learning. Moreover, the extraordinary value theory is utilized as a novelty finder procedure to recognize outliers, since the semi managed learning process is delicate to them. The learning machine is refreshed effectively. It can be utilized for different classes. Predominant properties are appeared for the proposed calculation as contrasted and an auto-encoder neural system.

The learning problem is again solved in [7]. It is effective and manages the communication efficiently. For prediction of label of unlabelled data the online graph based semi supervised learning algorithm is used. The data is arriving from clients. The output of this step is further used for the training of linear classifier. On client side, data is prioritized based on active learning metric technique. The communication is reduced by using the weight vector. This vector is sending from server to client periodically.

This paper [8] concentrated on the segmentation and quantitative assessment of the hydrocephalus from CT datasets. The proposed method works in three main stages: Firstly the extraction of the intracranial brain using traditional threshold. Second is the segmentation of hydrocephalus using algorithm depends on min-cut max-flow theorem and last is the HBR ratio calculation for volumetric dimensions for segmented areas. The min-cut max-flow algorithm is used to extract the hydrocephalus area gave extremely better results than the region growing techniques which are used previously.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

System developed online semi supervise learning and segmentation from ultra sound data. Automated left ventricle (LV) segmentation method and online learning segmentation algorithm can be used. It can be work on the problems such as image annotated databases and High time complexity to read all metadata from image objects in [9].

System designed a classic max-flow min-cut algorithm generalization that can be decreases the false data injection and attacks on power system. When multi node data transmission is performed the problems occurring like as no standard path selection mechanism and Energy wasted in [10].

III. SYSTEM ARCHITECTURE

Initially system read the dataset 1 and performs feature selection on the dataset. The algorithm used for feature selection is relief algorithm. Then classification of the dataset is performed using online semi supervised learning with graph cuts and max-flow algorithm as mentioned in ref1. The min-cut used is min-cut 3 which mean that there are 3 nearest neighbors to which an unlabeled point is connected while constructing graph. After that another batch of data is read. After that online semi supervised learning with graph cuts are performed. The output of semi supervised learning process is used to generate the graph. The graph is generated; Bernoulli's Random Variable Decision is applied to the graph. Classifier is updated. Innovations of the system are to design the system for Online Semi Supervised Learning algorithm and feature selection using relief algorithm.

In proposed system, for improving the efficiency of the system we are using cosine similarity for finding similarity which is best then the Euclidean distance used in base system.

We also made use of,

a. Feature selection for reducing features/ extracting features
b. Cosine similarity
c. Bernoulli's Random Variable

System proposes the online semi supervised learning technique with Max Flow and Min cut Algorithms.

a. For maximizing time efficiency
b. For minimizing memory utilization.
c. Maintaining accuracy

Proposed System Overview

Steps of proposed system:

1. Read Dataset
2. Preprocess data (convert to numeric format)
3. Feature selection using relief algorithm.
4. Create Batches of processed data.
5. Start first batch.
6. Calculate cosine similarity distance for un-labeled data.
7. Construct Graph using the Cosine similarity
8. Apply Mincut on Graph.
9. Update new classifiers.
10. Assign the label to the instances from the data which are connected in graph.
11. Read next batch
12. Apply cycle cancellation on graph for point out of sliding window.
13. Go to step 6.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

14. End.

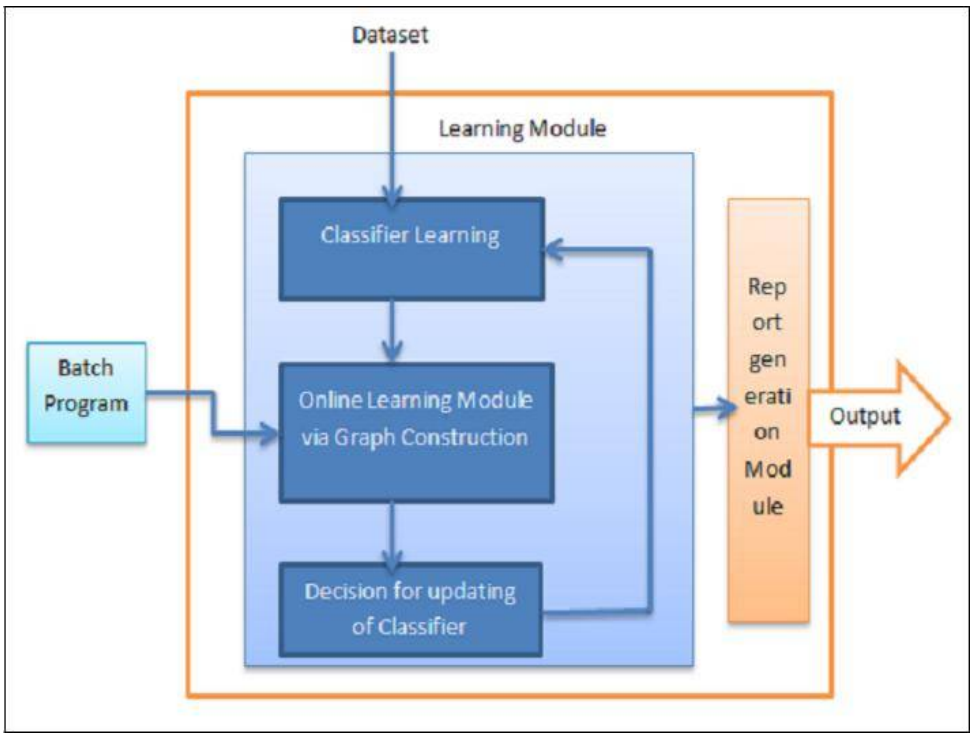

Fig. 1 System Architecture

IV. EXPERIMENTAL RESULTS

4.1 Experimental Setup

The system is built using Java framework on Windows operating system. The Net-beans IDE is used as a development tool. The system does not require any specific hardware to run. Application can run on any standard machine.

4.2 Dataset Used

We use KDD dataset for evaluation of proposed system. This dataset has 41 attributes. This dataset has nearly 30000 records. The network intrusion attacks can be classified in to two classes such as normal class and anomaly. After feature selection, it is observed that 23 relevant features in this dataset can be used to achieve good classification accuracy.

4.3 Evaluation Result

Figure 2 represent the graphical comparison of existing and proposed system on the basis of implementation time. Figure 3 represent the graphical comparison of existing and proposed system on the basis of their memory requirement.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

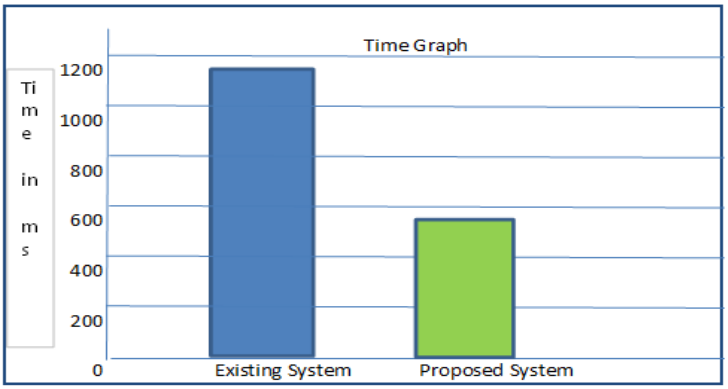

Fig. 2 Time Graph

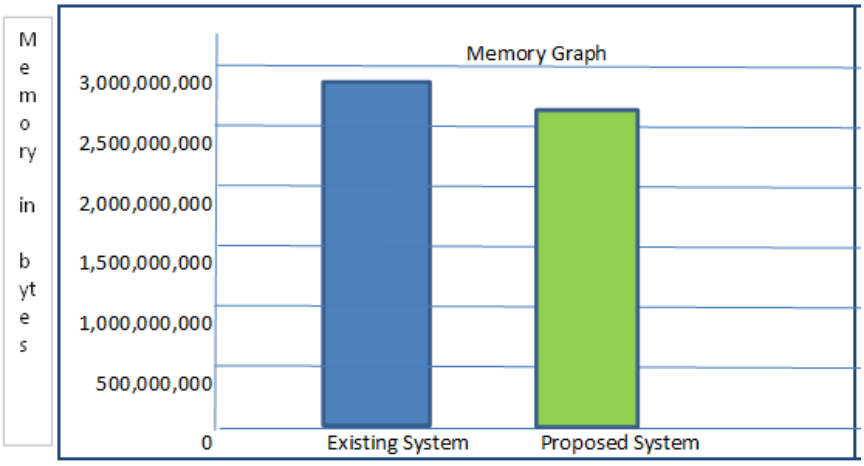

Fig. 3. Memory Graph

V. CONCLUSION

This paper explains the importance of semi-supervised scheme. The semi-supervised learning approach is has shown that the unlabeled data can be exploited for the sake of learning purpose. As the labeled data is very expensive, the semi-supervised learning has gain importance over other scheme especially in online applications where the amount of labeled data is less. Labeling the unlabeled data is very time consuming process hence Semi supervised algorithms are used for online learning. It needs less human efforts with better performance. This paper proves that the Combination of Max flow algorithm with Min Cut improves the accuracy of learning on online data streams. Also to improve the performance, cosine similarity and feature learning techniques are used. They improve the accuracy and time efficiency respectively. The performance of the system is tested on KDD dataset. Evaluation results prove the efficiency of system.

In online streams the data-stream grows speedily hence it requires a learning system which can cope up with this speed. Proposed algorithm provides a best way for fast and efficient serial computation. However the large amount of parallel processing of modules can be still achieved. The modules such as network construction, path search, network updating can be processed in parallel. In future, the system can also be incorporated with Hadoop environment so as to achieve parallel processing of tasks.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

VI. ACKNOWLEDGEMENT

The authors would like to thank the researchers as well as publishers for making their resources available. We are thankful of teachers for their valuable guidance. We are thankful to SavitribaiPhule Pune University. We are also thankful to the reviewer for their valuable suggestions.

REFERENCES

[1] Zhu, Lei, et al. "Incremental and decremental max-flow for online semi-supervised learning." IEEE Transactions n Knowledge and Data Engineering 28.8 (2016): 2115-2127.
[2] Mallapragada, Pavan Kumar, et al. "Semiboost: Boosting for semi-supervised learning." IEEE transactions on pattern analysis and machine intelligence 31.11 (2009): 2000-2014.
[3] Ye Tao, R. Triebel and D. Cremers, "Semi-supervised online learning for efficient classification of objects in 3D data streams" 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, 2015, pp. 2904-2910.
[4] M. Pratama, G. Zhang, M. J. Er and S. Anavatti, "An Incremental Type-2 Meta-Cognitive Extreme Learning Machine" in IEEE Transactions on Cybernetics, vol. 47, no. 2, pp. 339-353, Feb. 2017.
[5] C. Tan, G. Ji and B. Zhao, "SLFE: A New Semi-supervised Local Feature Extraction Algorithm," 2015 Third International Conference on Advanced Cloud and Big Data, Yangzhou, 2015, pp. 304-310.
[6] H. Al-Behadili, A. Grumpe, C. Dopp and C. Whler, "Extreme learning machine based novelty detection for incremental semi-supervised learning" 2015 Third International Conference on Image Information Processing (ICIIP), Waknaghat, 2015, pp. 230-235.
[7] H. Xiao, S. D. Lin, M. Y. Yeh, P. B. Gibbons and C. Eckert, "Learning better while sending less: Communication-efficient online semi-supervised learning in client-server settings" 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, 2015, pp. 1-10.
[8] Tomasz Weglinski, Anna Fabijanska, "Min-Cut/max-flow segmentation of hydrocephalus in children from CT datasets", International Conference on Signals and Electronic Systems (ICSES), 24 December 2012.
[9] Gustavo Carneiro, Jacinto C. Nascimento, "Incremental online semi-supervised learning for segmenting the left ventricle of the heart from ultrasound data", IEEE International Conference on Computer Vision (ICCV), 2011.
[10] Oliver Kosut"Max-Flow Min-Cut for Power System Security Index Computation", Sensor Array and Multichannel Signal Processing Workshop (SAM), 2014 IEEE 8th..