Review on Semi-Supervised Learning

Ms. Sujata Gawade[1], Prof. Vina M. Lomte[2]
[1, 2] Department of Computer Engineering
[1, 2] RMD Sinhagad Institute of technology, Warje Campus, Pune, Savitribai Phule, Pune University

Abstract- *in a common machine learning methods to classification, one can only make use of a labeled set from training the classifier. The problem with the labeled instances is those can be hard, expensive or may be very much time consuming to get because of the need the help of human annotators. Sometimes unlabeled data can be easy to get, but there have been some methods to utilize them. Semi-supervised learning solves these issues by making use of the huge size of unlabeled data, mixed with labeled data for creating the better classifiers. Due to the semi-supervised learning needs minimum human effort as well as provides greater accuracy. This survey presents some previous work done related to online supervised learning.*

Keywords- Labeled data, unlabeled data, semi-supervised learning, classifier.

I. INTRODUCTION

Semi supervised learning is a learning example related to the study of natural systems and computers like humans learn with the labeled as well as unlabeled data. Commonly, learning is analyzed in unsupervised paradigm such as clustering, outlier detection etc. where the data present is unlabeled or in the supervised paradigm such as classification, regression in which all data is labeled. The aim of the semi supervised learning is to know how the consolidation of labeled as well as unlabeled data can be modified the learning behavior as well as design a system which makes use of both of the methods. Semi supervised learning is manly used in machine learning as well as data mining due to its ability to use available unlabeled data to improve supervised learning tasks when the labeled data is scarce or expensive. Semi-supervised learning also shows potential as a quantitative tool to understand human category learning, where most of the input is self-evidently unlabeled.

Recent work on semi-administered learning accept there are two classes, and every class has a Gaussian distribution. This adds up to accepting the total information originates from a mixture model. With huge size of unlabeled data, the consolidated component can be related to the expectation-maximization (EM) algorithm. One needs just a single labeled example for each part to completely determine the mixture model. This model has been effectively connected to text categorization.

A variation is self-training: A classifier is initially trained with the labeled information. It is then used to arrange the unlabeled information. The most certain unlabeled points, together with their predicted labels, are added to the training set. The classifier is re-trained and the system repeated. Take note of the classifier utilizes its own particular forecasts to show itself. This is a "hard" form of the mixture model and EM algorithm. The strategy is likewise known as self-teaching, or bootstrapping1 in some exploration communities. One can envision that a classification mix-up can strengthen itself.

Both techniques have been utilized since long time back. They stay well known as a result of their applied and algorithmic simplicity.

Co-training lessens the mix-up strengthening risk of self-training. This late technique accept that the components of a thing can be part into two subsets. Every sub feature set is adequate to prepare a decent classifier; and the two sets are restrictively autonomous given the class. At first two classifiers are prepared with the marked information, one on every sub-feature set. Every classifier then iteratively classifies the unlabeled information, and instructs the other classifier with its predictions.

With the rising prominence of support vector machines (SVMs), transductive SVMs develop as an expansion to standard SVMs for semi-supervised learning. Transductive SVMs discover a labeling for all the unlabeled information, and an isolating hyper plane, with the end goal that most extreme edge is accomplished on both the labeled information and the unlabeled information. Naturally unlabeled information manages the choice limit far from dense regions.

As of late graph-based semi-supervised learning techniques have pulled in extraordinary consideration. Graph based strategies begin with a graph where the nodes are the marked and unlabeled information data points, and (weighted) edges reflect the similarity of nodes. The presumption is that nodes associated by a huge weight edge have a tendency to have a similar label, and label can propagation all through the graph. Graph-based strategies appreciate decent properties from spectral graph theory.

In this survey, Section II gives the Literature review for Semi-Supervised Learning systems and also list there pros and cons.

II. LITERATURE REVIEW

This paper [1] proposes, an online max-flow algorithm for the semi-supervised learning from data streams. In this firstly implement an online sample network whose pair-wised sample distance matrix is being updated 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. For learning from data streams with both labeled and unlabelled samples, an incremental detrimental max-flow based online semi-supervised learning system is proposed. For classification, compute min-cut over current max-flow, so that minimized number of similar sample pairs are classified into distinct classes. Empirical evaluation on real-world data reveals that the algorithm outperforms state-of-the-art stream classification algorithms.

This paper [2] presents, a boosting framework for semi-supervised learning, termed as Semi Boost. The strength of SemiBoost lies in its ability to improve the performance of any given base classifier. The empirical study on 21 different datasets shows that the proposed framework is effective for improving the performance of several supervised learning algorithms given a large number of unlabeled examples the results shows that the feasibility of this approach and the superior performance of SemiBoost compared to the state-of-the-art semi-supervised learning algorithms.

This paper [3] Based on Max-flow Min-cut Theorem, the method is used to find out the bottleneck of traffic network. According to the result of identification, the optimization of traffic network is made. The initial traffic network and the optimization is simulated by ExSpect software. The result of simulation demonstrates that the method of traffic network bottleneck identification based on Max-flow Min-cut Theorem can find out the traffic network bottleneck efficaciously.

This paper [4] proposed, a fair association scheme between clients and APs in WiFi network, exploiting the hybrid nature of the recent WLAN architecture. It show that such an association outperforms RSSI based schemes in several scenarios, while remaining practical and scalable for wide-scale deployment. It combines result from traditional graph theory with the emerging opportunities in wireless setting to target the pressing problem of association control. Initial results are promising both in terms of admittance and fairness.

This paper [5] Proposed, a Smart Association Control (SAC) protocol which consists of two algorithms: Fair Bandwidth Allocation (FBA) and Association for Maximum Throughput (AMT) algorithm. It combines results from traditional graph theory with the emerging opportunities in wireless setting to target the pressing problem of association control. SAC improves client association techniques in wireless networks by exploiting the wired backbone among WiFi APs. The key idea is to share local information from multiple APs, model it as a max- flow problem, and derive the optimal client-to-AP assignment. Simulation results demonstrate that such a technique can improve over purely distributed association schemes, resulting in higher fairness, better load balancing properties, and even some robustness to client mobility.

This paper [6] presents, a novel multi-output co-regularized learning algorithm (MOCA) and demonstrate its application on a real world problem in oral health domain an algorithm and a learning framework that is naturally suitable for the analysis of large scale, partially labeled metagenome datasets. It propose an online multi-output algorithm that learns by sequentially co-regularizing prediction functions on unlabeled data points and provides improved performance in comparison to several supervised methods. Evaluate predictive performance of the proposed methods on NIH Human Microbiome Project dataset. The proposed method outperforms several supervised regression techniques as well as leads to notable computational benefits when training the predictive model. In our semi-supervised modeling approach use the fact that interactions among different microbial species in various niches of human body are known to exist. In comparison to supervised approaches our methods leads to better predictive performance both on standard benchmark datasets from UCI repository and a real world HMP dataset.

This paper [7] presented a weight estimation algorithm (WEA) for determining classifier-voting weights when concept drift is present in incremental learning scenarios. WEA is an incremental ensemble based algorithm that uses labeled and to build classifiers and unlabeled data to aid in the calculation of the classifier voting weights before the data are classified. This contribution, describe an ensemble of classifiers based approach that takes advantage of both labeled and unlabeled data in addressing concept drift: available labeled data are used to generate classifiers, whose voting weights are determined based on the distances between Gaussian mixture model components trained on both labeled and unlabeled data in a drifting environment.

This paper [8] presents a new formulation of the voxel occupancy task. The new formula tion never computes a

silhouette, so it can handle noise in the original images as a well as situations where the object and background are of similar colors. The new formulation also naturally incorporates spatial smoothness which can improve the final results. An algorithm is presented which is based on graph cuts that can efficiently determine the 3D shape with lowest cost the smoothest shape which is consistent with the observations.

This paper [9] describes a new technique for general purpose interactive segmentation of N-dimensional images. The user marks certain pixels as "object" or "background" to provide hard constraints for segmentation. Additional soft constraints incorporate both boundary and region information. Graph cuts are used to find the globally optimal segmentation of the N-dimensional image.

This paper [10] uses a semi-supervised method to train classifiers to learn models for the foreground and background of an environment. Then, it uses the learned models as a way to bootstrap the overall system. A separate model is constructed to detect the changes in the background. It is then integrated together with audio prediction models to decide on the final background or foreground determination. The issue is the ratio of unlabeled data is on higher side.

III. PROPOSED APPROACH

Semi-supervised learning is initially motivated by its practical value in learning faster, better, and cheaper. In many real world applications, it is relatively easy to acquire a large amount of unlabeled data. For example, documents can be crawled from the Web, images can be obtained from surveillance cameras, and speech can be collected from broadcast. However, their corresponding labels for the prediction task, such as sentiment orientation, intrusion detection, and phonetic transcript, often requires slow human annotation and expensive laboratory experiments. This labeling bottleneck results in a scarce of labeled data and a surplus of unlabeled data. Therefore, being able to utilize the surplus unlabeled data is desirable.

Recently, semi-supervised learning also finds applications in cognitive psychology as a computational model for human learning. In human categorization and concept forming, the environment provides unsupervised data in addition to labeled data from a teacher. There is evidence that human beings can combine labeled and unlabeled data to facilitate learning.

In proposed system two datasets are used. First data set is used for classifier learning and second classifier is use as a test dataset as an input to the system. After that the graphs are generated by making use of labeled and unlabeled data. Then by conducting min cut separation labels are assigned and graphs are updated. At last classification is performed.

IV. RESEARCH GAP

This section describes the research gap of recently developed systems. This paper studied the various techniques that contributes into the field of semi supervised learning and identifies the respective limitations. This research gap provides the further improvements needed to this field.

- Improves the overall throughput of the entire network
- It should resilient to client mobility.
- Augmenting the system with sophisticated channel and traffic models.

V.CONCLUSION AND FUTURE SCOPE

This paper analyses various techniques used for Semi-Supervised Learning. Also given the advantages and drawbacks present in the different studies performed by various researchers. To deal with drawbacks in present systems we presented an idea of the new system..

REFERENCES

[1] Zhu, Lei, et al. "Incremental and Decremental Max-flow for Online Semi-supervised Learning."

[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] Dong, Shengwu, and Yi Zhang. "Research on method of traffic network bottleneck identification based on max-flow min-cut theorem." Transportation, Mechanical, and Electrical Engineering (TMEE), 2011 International Conference on. IEEE, 2011.

[4] Dandapat, Sourav Kumar, et al. "Fair bandwidth allocation in wireless mobile environment using max-flow." 2010 International Conference on High Performance Computing. IEEE, 2010.

[5] Dandapat, Sourav Kumar, et al. "Smart association control in wireless mobile environment using max-flow." IEEE Transactions on Network and Service Management 9.1 (2012): 73-86.

[6] Imangaliyev, Sultan, et al. "Online semi-supervised learning: algorithm and application in metagenomics." Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on. IEEE, 2013.

[7] Ditzler, Gregory, and Robi Polikar. "Semi-supervised learning in nonstationary environments." Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE, 2011.

[8] Selina Chu, Shrikanth Narayanan, C.-C. Jay Kuo, "A Semi-supervised learning for audio background detection", IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.

[9] Y. Y. Boykov, M.-P. Jolly"Interactive Graph Cuts for Optimal Boundary Region Segmentation of Objects in N-D Images", Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001

[10] Dan Snow, Paul Viola and Ramin Zabih, "Exact Voxel Occupancy with Graph Cuts", IEEE Conference on Computer Vision and Pattern Recognition, June 2000.