Learning a Propagable Graph for Classification under the Scenario with Uncertain Labels

[1]Mr Karan P. Bhatt, [2]Ms Ompriya V. Kale
[1]Student of M.E., [2]Assistant Professor
[1]P.G. C.E. Department,
[1]L.J.I.E.T., Ahmedabad, India

Abstract **- In this paper, we present the classification of uncertain labels using learning by propagability in graph. The entire leaning methodology is based on the data labels and optimality of feature representation that can create a harmonic system. Here data labels are invariant regarding the propagation on the similarity graph constructed based on the optimal feature representation. Using this idea, we can perform classification. This approach deals with multiclass classification problems. Dealing with uncertain labels is the crucial problem in this approach. Using graph-based semi-supervised learning, we can implement the classification procedure for uncertain labels data sets. Different from previous graph-based methods that are based on discriminative models, our method is essentially a generative model in that the class conditional probabilities are estimated by graph propagation and the class priors are estimated by linear regression.**

IndexTerms **- Semi-supervised learning, graph-based learning, propagability, graph harmoniousness, uncertainty, feature extraction, multiclass classification**

I. INTRODUCTION

Data mining is the technique to extract features from raw data. It is used as a basic of knowledge discovery of data. Data mining is commonly used in areas like genetics, education, finance, bio-informatics, pattern matching, genetic science and generation of probabilistic models.

There are several drawbacks of data mining. In some applications, such as facial images, satellite images, human body internal structure images, feature representation is extremely high dimensional. These high dimensional extracted features are sometimes nothing but noisy and redundant dimensions. Therefore, the transformation of high dimensional data into low dimensional data with more informative feature space is highly suitable.[1] Data mining is unable to provide sufficient labeled training samples. Applications like human face labeling have extremely expensive data. To get sufficient labeled training samples, graph-based learning mechanism is one of the efficient methodologies.

Propagability of a graph is one of the crucial factors which have to deploy during learning. Graph mining provides better results in certain labels whereas uncertain labels treatment is difficult. Therefore, using semi-supervised learning mechanisms on propagable graph and concept of belief and certainty factors, one can perform tasks like classification, regression, pattern extraction and clustering. Uncertain datasets are taking place widely in today's day-to-day life. Therefore, extraction is needed for it.

For label propagation, first we have to construct a graph using optimal low-dimensional feature space. Generally, semi-supervised learning algorithms provides two-class classification problem. For propagable graph, multiclass classification is more appropriate and even much more efficient.[1] The iterative process should be constructed so we can perform multiclass classification directly. Labeled data and unlabeled data, both must be trained and sum-to-one constraints are always be true.

Propagation coefficient is the crucial factor to get predicted class probability. Data with uncertain labels can be classified with the concept of belief functions.[3] Each training pattern is associated with a basic belief assignment, showing partial knowledge of its actual class. Uncertain labels in a propagable graph are classified and trained through learning approach with the help of concept of information theory.

General methods for classification of certain labels through graph-based learning cannot suitable because every time probability of labels is changed and it is not robust.[1] To get appropriate result for uncertain labels, we have to apply modification in current working algorithms for certain labels. Uncertain labels are one of the issues in the data sets like breast cancer.

II. THEORITICAL BACKGROUND AND LITERATURE SURVEY:

A. THEORY OF MACHINE LEARNING:

Machine learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Machine learning on the development of computer programs that can teach themselves to grow and change when exposed to new data.[1]

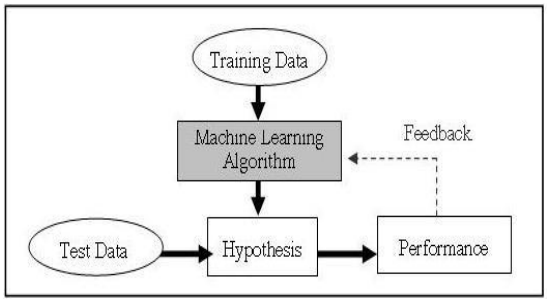

Fig. 2.1[5]

When we want highest level of abstraction, we can think of machine learning as a set of methodology and appropriate guide that mark to infer patterns and extract new one from a record of the observations. In machine learning, the learning occurs by extracting as much information from the data as possible through algorithms that track the basic structure of the data and distinguish the actual data. After they have found the actual data, or pattern, the algorithms simply decide that everything else that is left over is useless data. Therefore, machine learning techniques are also referred to as pattern recognition tool.

As shown in Fig. 2.1, test data are given as input to formulate hypothesis first. To train the data sets, we gives training data as input to machine learning algorithm. Outcome of this algorithm is also become input to formulate hypothesis. Using hypothesis, we can evaluate performance and the feedback of this performance again provided to machine learning algorithms. There are mainly two types of learning:

1) Supervised Learning

Learning mechanism of inferring a function from labeled training data

2) Unsupervised Learning

Learning mechanism of inferring a function from unlabeled training data

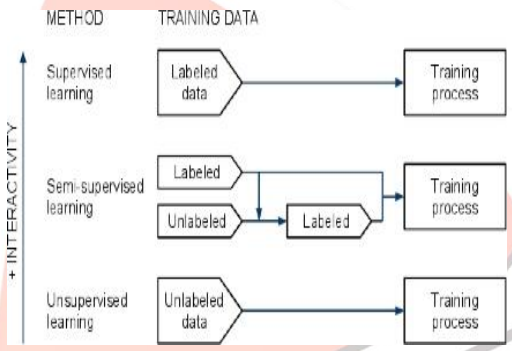

Fig. 2.2[10]

The third approach is semi-supervised learning which is actually a type of supervised learning which deals with unlabeled data also.

B. GRAPH BASED LEARNING:

The basic idea behind graph-based learning is as follow:

Step 1:
Construct a graph connecting similar data points

Step 2:
Let the hidden/observed labels be random variables on the nodes of this graph (i.e. the graph is an MRI)

Step 3:
Similar data points have similar labels

Step 4:
Information “propagates” from labeled data points

Step 5:
Graph encodes intuition

Step 6:
Apply semi-supervised learning algorithm with appropriate learning rule [9]

Here two objects are similar if they are neighbors. In the field of social-networks, for example, each individual is represented as a node in a graph, with a link between two nodes if the individuals are friends. Here from a graph one might wish to identify communities of closely linked friends.[3] Discovering such groupings contrasts with graph partitioning in which each node is assigned to only one of a set of sub graphs for which a typical criterion is that each sub graph should be roughly of the same size and that there are few connections between the sub graphs.[1]

C. NEED OF FEATURE EXTRACTION:

Feature extraction in graph-based learning is formulated by learning rules. First of all we have to perform transformation from high dimension feature space to low dimension feature space. This transformation provides better relationship between data, shown in Fig.2.3.

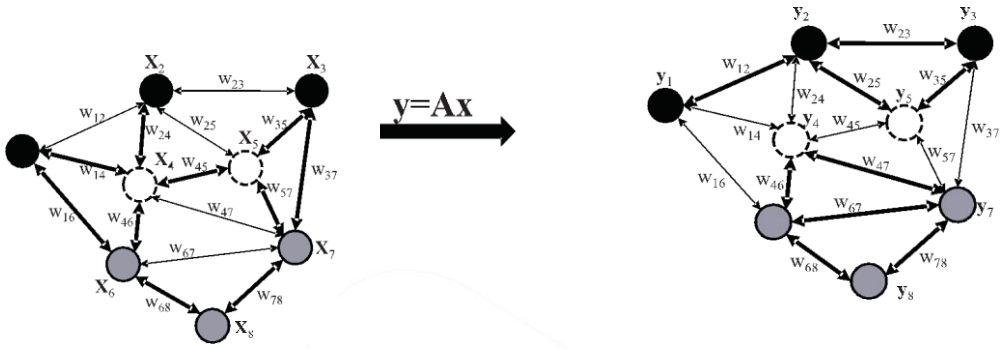

Fig. 2.3[1]

After transformation, we have to characterize harmoniousness of the graph by mathematical formulation. It provides classification capability of the derived feature representation. Here we can provide ranking of a node procedure too. It gives great improvement in output.

D. LEARNING WITH UNCERTAIN DATA:

The method is designed to tackle the binary classification problem under the condition that the number of labeled data points is extremely small and the two classes are highly imbalanced. It starts with only one positive seed given by the contest organizer.[12]

We randomly pick additional unlabeled data points and treat them as “negative” seeds based on the fact that the positive label is rare across all datasets. A classifier is trained using the “labeled” data points and then is used to predict the unlabeled dataset. We take the final result to be the average of n stochastic iterations.

Procedure: Stochastic semi-supervised learning process

Given one positive label, N unlabeled examples:

1. For dataset A, C and D
2. Set i = 1
3. Randomly pick one example from N unlabeled examples as “negative" example
4. Use the positive label and the \negative" label as initial seeds, do k-means clustering on whole dataset with no. of cluster = 2
5. Label cluster where positive seed sits as positive, another one as negative
6. Save cluster membership of each example fi(c) where f(c = positive) = 1; f(c = negative) = 0.
7. Increase i by 1. If i < 100 return to step3
8. Calculate final predicted score for each example using $1/M \Sigma_{i=1..M} = 1$ fi(c) where M =100.
9. For dataset B, E and F
10. Set i = 1
11. Randomly pick 20 examples from N unlabeled examples as “negative" examples
12. Use the one positive label and the 20 \negative" labels as training set, build a logistic regression model
13. Score the whole dataset, save score for each example fi
14. Increase i by 1. If i < 100 return to step 11.
15. Calculate average score for each example using $1/M \Sigma_{i=1..M} = 1$ fi where M = 100.
16. Label highest 1% of score as positive examples, lowest 1% of score as negative examples, rebuild the logistic regression model (self-training).
17. Calculate final score for each example using above logistic regression model. [8]

The above semi-supervised learning approach is used when the number of available labels is extremely small. When the amount of the labeled data becomes large, we tend to use Gradient Boosting Decision Tree (TreeNet) as our classifer to generate prediction score.

E. GRAPH HARMONIOUSNESS:

Graph harmoniousness is necessary and basic condition in any graph mining techniques. A connected labeled graph with n edges in which all vertices labeled with distinct integers (mod n) so that the sums of the pairs of numbers at the ends of each edge are considered as a distinct edge (mod n). The ladder graph, fan, wheel graph, Petersen graph, tetrahedral graph are all harmonious. The necessary conditions for the existence of odd harmonious labeling of graph are obtained. A cycle C_n is odd harmonious if and only if $n \equiv 0 \pmod 4$. A complete graph K_n is odd harmonious if and only if $n=2$. A complete k-partite graph $K(n_1, n_2, \ldots, n_k)$ is odd harmonious if and only if $k=2$. A windmill graph K_n^t is odd harmonious if and only if $n=2$.[11]

III. PROPOSED FLOW CHART

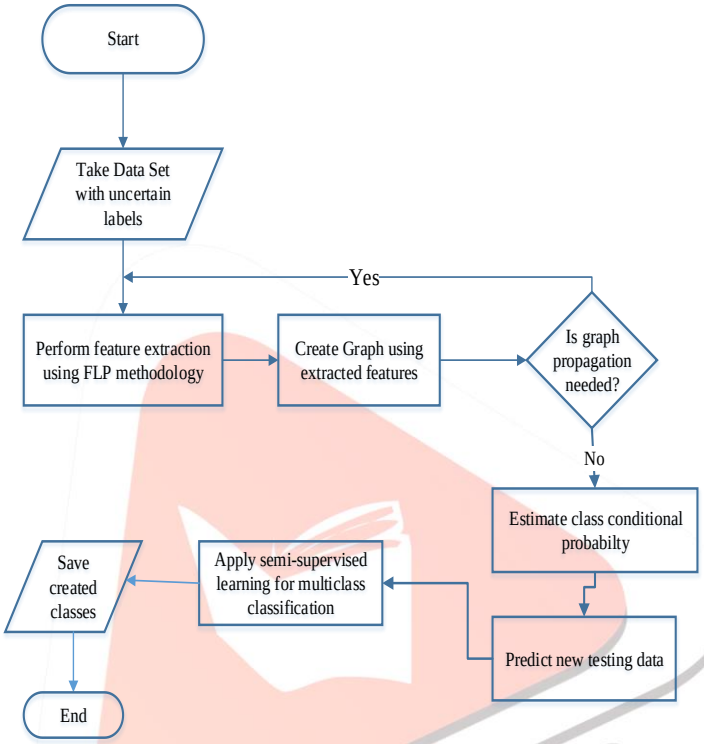

Fig. 3.1

IV. RESULT ANALYSIS

Data Set: Breast Cancer Wisconsin (Original) Data Set

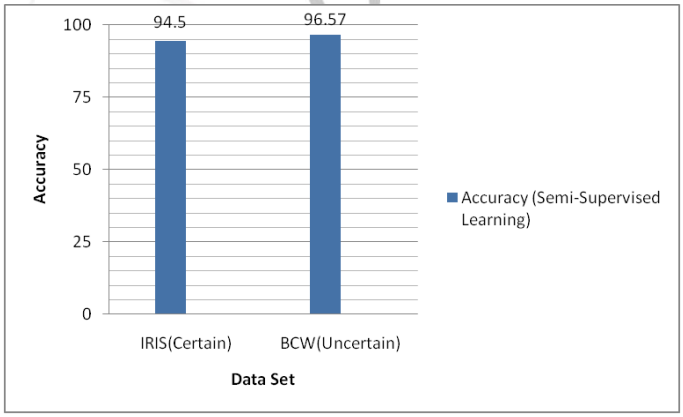

Fig. 4.1 Result analyses with semi-supervised learning

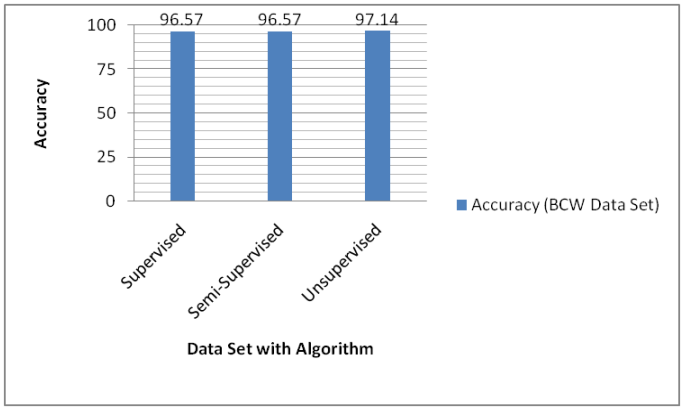

Figure 4.2 Result comparisons of all learning methods

V. CONCLUSION

Graph-based learning provides a set of powerful techniques for data analysis and predictive analytics. Learning a propagable graph provides a novel approach of feature extraction which is more efficient. Classification of uncertain labels is one of the crucial tasks which is performed by semi-supervised learning with the help of certainty factors. In future we can perform regression on the same uncertain labels.

REFERENCES

[1]Bingbing Ni, Shuicheng Yan and Ashraf A. Kassim, "Learning a Propagable Graph for Semisupervised Learning: Classification and Regression", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[2]Chang-Dong Wang, *Jian-Huang Lai and* Jun-Yong Zhu, "Graph-Based Multiprototype Competitive Learning and Its Applications", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PARTC APPLICATIONS AND REVIEWS, VOL. 42, NO. 6, NOVEMBER 2012

[3]Adriana Prado, Marc Plantevit, Ce´line Robardet, and Jean-Franc¸ois Boulicaut, "Mining Graph Topological Patterns: Finding Covariations among Vertex Descriptors", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 9, SEPTEMBER 2013

[4]Senjuti Basu Roy, Tina Eliassi-Rad and Spiros Papadimitriou, "Fast Best-Effort Search on Graphs with Multiple Attributes", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 2014

[5]Paolo Napoletano, Francesco Colace, Massimo De Santo, Luca Greco, "Text classification using a graph of terms", 2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems

[6]U Kang and Christos Faloutsos, "Big Graph Mining: Algorithms and Discoveries", SIGKDD Explorations Volume 14, Issue 2, March 2013

[7] Joseph A. Gallian, "A Dynamic Survey of Graph Labeling" published by The Electronic Journal Of Combinatorics, Edition 16

[8]Sumeet Dua and Xian Du, "Data Mining and Machine Learning in Cybersecurity", Auerbach Publications, May 2011

[9]Aggarwal, Charu C., Wang, Haixun (Eds.), "Managing and Mining Graph Data", Kluwer Academic Publishers Boston/Dordrecht/London, July 2010

[10]http://www.intechopen.com/books/vehicular-technologies-deployment-and-applications/smart-vehicles-technologies-and-main-applications-in-vehicular-ad-hoc-networks

[11]http://www.enterprisetech.com/2014/02/11/netflix-speeds-machine learning-amazon-gpus/

[12]http://cs229.stanford.edu/materials.html