Inferring Age and Gender via Graph-Based algorithms in Big Data Platforms

Subtitle:-Survey paper on the label propagation or node classification methods that can be applied to the graph to make label predictions.

Sneha Gunda

Graduate Student

University of Washington, Tacoma

E-mail:- snehag@uw.edu

1. Abstract:-

In this survey paper I would like to study the label propagation or node classification algorithms that is to be applied to the graph to derive unknown labels of nodes that will predict the labels of a node such as age and gender. Label propagation is one of the effective learning method used in various applications. In Label propagation, we construct a graph where a nodes represent a data point and an edge that includes weight representing the similarity between the data points i.e. there will be an edge between two data points if and only if they have at least one feature in common. After construction of the graph, we will make predictions based upon the graph.

Keywords:-

Label Propagation, attribute, Node propagation, label, Random Walk Methods, Iterative Classification Methods.

2. Introduction:-

2.1 Motivation:- Data can be divided into labeled and unlabeled data. Generally labeled data is very essential for a supervised learning. There is a lot of raw data but labeled and organized data may be in small quantities and limited [2]. It will be very easy to process the data if it is in an organized manner so, conversion of the given raw or unlabeled data into a graph with nodes where each node is labeled and edges where each edge has a weight is very essential.

In some cases, users of a social network only reveal some information about themselves and hide the rest so, a graph may or may not have a label and the label can be in different forms such as a binary number, a numeric value, or a range of numbers or a textual data and label propagation is used to identify unknown label [1]. We have to make sure that all the nodes in a graph are labeled. The node classification algorithm is applied only if there is only one value for the label of a node.

2.2 Problem statement:-A graph with a subset of nodes labeled and others unlabeled and a set of labels for the entire graph will be considered as input and we will apply the label propagation or node classification algorithms to predict the label of nodes for which it is unknown. There are many label propagation algorithms that can be analyzed in top-to-down approach or bottom-up approach [2]. I will be analyzing labeling algorithms such as “random walk” or “iterative classification method” supervised learning algorithms to find the label or attribute of new nodes [1].

2.3 Approach:-

In a social network, large part of information about individuals and their activities can be modeled as labels for a specific user. They can also be converted to represent the nodes of a graph. The problem of knowing the labels for individuals is a key element for any application. So, in my survey paper, my main aim is to analyze the node classification algorithms and apply it to the problem statement and predict the label of an unknown user.

There are two approaches to solve the node classification problem:-

2.3.1 Iterative methods:- These methods use the neighborhood information to generate features that are used to learn local classifiers which are used to label the nodes. This method is applied iteratively until there are no more unlabeled nodes and the label of the nodes doesn’t change in the next iterations [1] [3].

If ‘V’ is a set of nodes in a graph then set of attributes for each node ‘$v_i \in V$ will be known which are included as a subset of “feature vector” for that specific node. The “link features” consists of the information or relation between the current and the neighboring nodes or it can also be defined as the frequency with which a label present on the current node also appears on the neighboring node. There can be more than one link features if there are directed edges between two nodes, the link feature also depends upon the degree of the nodes.

Ø = Feature vector matrix for all nodes in ‘V’

$Ø_i$ = Feature vector for node v_i

$Ø_l$ = Feature vectors for labeled nodes

$Ø_u$ = Feature vectors for unlabeled nodes

V = Set of all nodes

V_l = Set of initially labeled nodes

V_u = Set of initially unlabeled nodes

Y = Set of all labels

Y_l = Set of labels on nodes in V_l

Y_u = Set of labels for nodes in V_u

Y_u^t = Set of new labels at ‘t’th iteration [1]

Algorithm:-

```
Iterative Classification Algorithm (ICA)
{
Train the initial classifier using $Ø_l$ and $Y_l$
Apply the trained initial classifier to $Ø_u$ to calculate $Y_u^1$ and the new feature vector
Repeat the process for ‘t’ iterations //At each iteration, the feature vector changes so a new feature vector is calculated
At ‘t’th iteration, a feature vector $Ø_t$ is obtained and is applied to calculate new labels $Y_u^t$.
Repeat the process until all the nodes are labeled and until no label changes in an iteration
}
```

In some cases, the stability may not be reached that is there is no guarantee that the label names remain constant so, in this cases we can choose a fraction denoting a set of labels that does not change.

2.3.2 Random walk based methods:- These are semi-supervised learning methods or transductive learning methods where a global labeling function over the graph is learnt. The walk is randomly performed over nodes to calculate the labels. This method uses the link structure for labeling the nodes. In this method, the probability of labeling a node belonging to the graph is equal to the total probability of a random walk starting from a particular node and ending at a labeled node. In this method, the graph is assumed to be label connected i.e. it is possible to reach an unlabeled node from a labeled node in many ways [1].

A transition matrix ‘P’ consisting the probabilities such that $\sum_j^n Pij = 1$ and $0 \leq P_{ij} \leq 1$ is defined if an element consists of value equal to ‘1’ then there will be zero probability of leaving the node(v_j) starting from v_i which acts as the termination condition for a random walk. The random walk algorithm converges to stationary labeling values at P^{α}.

$\sum_j^n Pij = 1$ and $0 \leq P_{ij} \leq 1$ → Transition matrix

$P_{ij}{}^{t}$ → Probability of reaching node v_j after ‘t’ steps

P^t → Transition matrix at time ‘t’

P^{α} → the probability matrix at $t=\alpha$

P_{ll} → Matrix with probabilities corresponding to the transactions from labeled node to a labeled node

P_{lu} → Matrix with probabilities corresponding to the transactions from labeled node to an unlabeled node

P_{ul} → Matrix with probabilities corresponding to the transactions from an unlabeled node to a labeled node

P_{uu} → Matrix with probabilities corresponding to the transactions from unlabeled node to an unlabeled node

The equation for classification using random walks is:

$Y' = P^{\alpha}Y$ where ‘Y’ is the matrix of probability distributions on the label set

Algorithm:-

```
Random walk formulation Algorithm

{

Choose an out-going edge from any node $v_{i.}$

If ($v_i$ is already labeled)

    {
      They are the absorbing state nodes so,
    no change in labels
    }

Else

   {
             The edge with probability
             proportional to the edge weight
             is chosen

   }
```

Build the transition matrix P = $\begin{pmatrix} \mathrm{Pll} & \mathrm{Plu} \\ \mathrm{Pul} & \mathrm{Puu} \end{pmatrix} = \begin{pmatrix} \mathrm{I} & 0 \\ \mathrm{Pul} & \mathrm{Puu} \end{pmatrix}$ //as on reaching labeled node the algorithm ends so the probability here (Pll) is ‘1’ and the labeled node can never become unlabeled so, the probability matrix (Plu) is always ‘0’.

Random walk algorithm terminates with all labeled nodes at P^{α}

$$\mathrm{P}^{\alpha} = \begin{pmatrix} \mathrm{I} & 0 \\ \mathrm{Pul}(\mathrm{I} - \mathrm{Puu})^{-1} & \mathrm{Puu}^{\alpha} \end{pmatrix}$$

Calculate the new labels using the P^{α} value as

Y’ = $\mathrm{Pul}(\mathrm{I} - \mathrm{Puu})^{-1}\mathrm{Yl}$

}

3. Experiment or Evaluation:-

The experimental design or the demo of the project will include running the code for the random walk algorithm for a drunken man who walks randomly on a grid. The algorithm will take the size of the grid, the coordinates of ‘x’ and ‘y’ as the starting points and moves forward randomly until one of the end of the grid is reached.

Schedule:-

I have divided the entire survey into 3 tasks that can be achieved in three weeks.

Task1:- Identify and study the published papers in node classification.

Task2:- Analyze the iterative classification and random walk algorithms that is to be applied to the graph.

Task3:- Implement the algorithm in some programming language.

4. Deliverables:-

My deliverables include finding a label propagation algorithm for a graph with nodes and edges constructed from the “Netlog” or “Facebook” data and implement a code which predicts the age and gender of a user and prepare presentation slides of my work.

5. Related work:-

Paper-1:- *Node classification in social networks*

Labeled nodes are used to extend information for unlabeled nodes so, for such a naming, the two methods such as iterative application of traditional classifiers to a graph and random walk method which transfers through the labeled nodes, node classification, graph labeling are described in this paper.

Paper-2:- *learning from labeled and unlabeled data with label propagation*

This paper mainly describes how a node’s label can be propagated from the adjacent nodes. Here, they describe iterative algorithms, label propagation in a dense and large network and how labels can be propagated from random walk and clamping algorithms. Description of the algorithm, how it leads to the solution and comparison with other algorithms like minimum spanning tree is provided.

Paper-3:- *Relation extraction using label propagation based semi-supervised learning:-*

This paper provides a description and an algorithm on how semi-supervised learning is performed on a graph. It presents a labeling function for a labeled node and also for the entire graph considering the nodes and edges of a graph. It also describes how labeling algorithm performs better than the other algorithms.

Paper-4:- *Labels vs. pairwise constraints: A unified view of label propagation and constrained spectral clustering*

The process of labeling a node can be done using two methods “label propagation” where unknown nodes are labeled and “constrained spectral clustering” where labels are converted into pairwise constraints (must-link and cannot-link). The paper describes that even though these two fields are separately developed, they are related to each other [4].

Paper-5:- *Balanced label propagation for partitioning massive graphs*

This paper introduces “Balanced label propagation” algorithm for partitioning massive graphs. It is an algorithm that can partition a massive graph or a graph with billions of edges. It combines the efficiency of label propagation with constrained optimization and its performance is evaluated on the social networks such as Facebook.

7. References:-

[1] Bhagat, S., Cormode, G., and Muthukrishnan, S. Node classification in social networks. Computing Research Repository (CoRR) abs/1101.3291 (2011). http://dimacs.rutgers.edu/~graham/pubs/papers/graphlabelchapter.pdf

[2] Zhu, X., Ghahramani Z., Learning from labeled and unlabeled data with label propagation. Carnegie Mellon University (CMU-CALD)-02-107 (June 2002). http://www.cs.cmu.edu/~zhuxj/pub/CMU-CALD-02-107.pdf

[3] Jinxiu Chen, Donghong Ji, Chew Lim Tan, Zhengyu Niu, Relation extraction using label propagation based semi-supervised learning. http://delivery.acm.org/10.1145/1230000/1220192/p129-chen.pdf?ip=71.227.178.99&id=1220192&acc=OPEN&key=BF13D071DEA4D3F3B0AA4BA89B4BCA5B&CFID=269788591&CFTOKEN=56202479&__acm__=1386702701_1e6843539da19bfcac5dbaef76edc2d5

[4] Wang, X., Qian, B., Davidson, I., Labels vs pairwise constraints: A unified view of label propagation and constrained spectral clustering, University of California, Davis.

[5] Ugander, J., Backstrom, L., Balanced Label Propagation for Partitioning Massive Graphs, Cornell University. https://people.cam.cornell.edu/~jugander/papers/wsdm13-blp.pdf

[6] Hassan, S., Mihalcea, R., Banea, C., Random-walk term weighting improved text classification, University of North Texas. http://www.cse.unt.edu/~rada/papers/hassan.ieee07.pdf