Expert Systems With Applications 45 (2016) 234–247

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Data heterogeneity consideration in semi-supervised learning

Bilzã Araújo [a,b,*], Liang Zhao [c]

[a] Institute of Humanities, Arts and Sciences, Federal University of Southern Bahia, BR-367, Km 10, CEP: 45810-000, Porto Seguro, Bahia, Brazil
[b] Department of Computer Science, Institute of Mathematics and Computer Science, University of São Paulo, Av. Trabalhador São-carlense, 400, Caixa Postal: 668, CEP: 13560-970, São Carlos, São Paulo, Brazil
[c] Department of Computation and Mathematics, School of Philosophy, Science and Literature in Ribeirão Preto, University of São Paulo, Av. Bandeirantes, 3900, CEP: 14090-901, Ribeirão Preto, São Paulo, Brazil

ARTICLE INFO

Keywords:
Semi-supervised learning
Graph construction
Complex networks
Representatives selection
Principal components analysis

ABSTRACT

In class (cluster) formation process of machine learning techniques, data instances are usually assumed to have equal relevance. However, it is frequently not true. Such a situation is more typical in semi-supervised learning since we have to understand the data structure of both labeled and unlabeled data at the same time. In this paper, we investigate the organizational heterogeneity of data in semi-supervised learning using graph representation. This is because graph is a natural choice to characterize relationship between any pair of nodes or any pair of groups of nodes, consequently, strategical location of each node or each group of nodes can be determined by graph measures. Specifically, two issues are addressed: (1) We propose an adaptive graph construction method, we call AdaRadius, considering the heterogeneity of local interacting structure among nodes. As a result, it presents several interesting properties, namely adaptability to data density variations, low dependency on parameters setting, and reasonable computational cost, for both pool based and incremental data. (2) Moreover, we present heuristic criteria for selecting representative data samples to be labeled. Experimental study shows that selective labeling usually gets better classification results than random labeling. To our knowledge, it still lacks investigation on both issues up to now, therefore, our approach presents an important step toward the data heterogeneity characterization not only in semi-supervised learning, but also in general machine learning.

1. Introduction

Semi-supervised learning (SSL) is known as a mid-term between unsupervised and supervised machine learning paradigms where both unlabeled and labeled data are taken into account in class or cluster formation and prediction process (Chapelle, Schlkopf, & Zien, 2006; Zhu, 2008). In real world applications, we usually have partial knowledge on a given dataset. For example, we certainly do not know every soccer players, but we know some famous ones; in a large scale social network, we usually just know some friends; in biological domain, we are far away to have a complete figure on all the protein functions, but we know the functions of some of them. Sometimes, although we have a complete or almost complete knowledge on a dataset, the labeling by hand is lengthy and cost, so, it is necessary to restrict the labeling scope. For these reasons, partially labeled datasets are often encountered. In this sense, supervised and unsupervised learning can be considered as extreme and special cases of semi-supervised learning. Up to now, many semi-supervised learning techniques have been developed, including generative models (Nigam, McCallum, Thrun, & Mitchell, 2000), clustering and labeling techniques (Wagstaff, Cardie, Rogers, & Schroedl, 2001), multi-training (Blum & Chawla, 2001; Zhou & Li, 2005), low-density separation models (Vapnik, 1998), and graph-based methods (Zhu, 2005). Among the above listed approaches, graph-based SSL has been triggered much attention. In this case, data is represented relationally (Belkin, Niyogi, & Sindhwani, 2006; Zhu, 2005). Each data instance is represented by a node and it is linked to other nodes according to a predefined affinity rule. The graph construction from vector-based data is equivalent to non-linear data dimensionality reduction or manifold learning (Belkin & Niyogi, 2003; Belkin et al., 2006; Roweis & Saul, 2000; Zhu, 2005). The label propagation, in turn, is similar to the transfer of beliefs on trust networks, epidemic spreading on contact networks, spreading of computer viruses on email networks and information spreading on social networks (Barrat, Barthlemy, & Vespignani, 2008; Barthélemy, Barrat, Pastor-Satorras, & Vespignani, 2004; Newman, Forrest, & Balthrop, 2002; Pastor-Satorras & Vespignani, 2001). Two well-known assumptions for the effectiveness of

* Corresponding author. Tel.: +55 73 32888400.
E-mail addresses: bmarques@gmail.com, bilza@ufsb.edu.br (B. Araújo), zhao@usp.br (L. Zhao).

http://dx.doi.org/10.1016/j.eswa.2015.09.026
0957-4174/

the graph-based SSL are the smoothness and clustering assumptions. Nearest nodes should be labeled with the same labels and nodes in the same modular structure should be labeled with similar labels. That is, class distributions should conform with graph structure, and vice-versa.

To our knowledge, all of the SSL techniques consider that all the data instances have equal relevance, i.e., the particular function of each data instance to the whole dataset is not taken into consideration. This assumption is frequently not true. For example, in the soccer player dataset, the famous stars, such as Pele and Maradona, are more representative than others; in a social network, the concept of "importance" is dynamic and it varies from one to another, i.e., each one may consider a different group of people to be more important than others (Borge-Holthoefer & Moreno, 2012; Borge-Holthoefer, Rivero, & Moreno, 2012); again in biological domain, some proteins are more decisive than others in a given living organism. Moreover, in some living organisms, some proteins can be even ignored (Jeong, Mason, Barabsi, & Oltvai, 2001; Mewes et al., 2004). These facts suggest that learning performance can be improved if the individual differences among data are considered. Recent results have shown that even clustering and smoothness assumptions of semi-supervised learning are satisfied, the label propagation performance may vary according to the labeled instances, which means that labeling a randomly chosen subset of data instances is not always a good strategy. Therefore, it is necessary to identify representative data instances in a given dataset and, consequently, provide guidelines for labeling process by human experts or by computer.

In this paper, we address the data heterogeneity issue in graph-based semi-supervised learning. We first present an adaptive graph construction method to transform the vector-based dataset into a graph taking into account the local/global structure of individual data instance. The local/global structure tells the distinct function played by each node and subsequently, indicates the proper subset of nodes, which it should be connected. Graph construction is a crucial process in graph-based learning, but it remains a challenging problem. Traditionally, a graph is constructed from a vector-based dataset, where each node represents a data instance and each node is connected to some other nodes using a predefined affinity rule, for example, each data instance is connected to the k most similar nodes, called k nearest neighbors (kNN) method; or each data instance is connected to all other instances within a certain distance ϵ, called ϵ-radius method; or the combination of them. In neither case, the data heterogeneity or the data local/global structure is considered. Another drawback of such kind of methods is that it is hard or even impossible to choose an affinity rule which is the most beneficial for all datasets (Carreira-perpiñán & Zemel, 2004; de Sousa, Rezende, & Batista, 2013; Jebara, Wang, & Chang, 2009; Rohban & Rabiee, 2012; Zhu, 2008). Here, we propose an adaptive graph construction method based on the Minimum Spanning Tree (MST). Roughly speaking, we compute the pairwise distance between each pair of data items and then, we find out the MST, which serves as a skeleton of the yielded graph. Over the MST, we estimate the coverage radius for each node through which we set up remaining links. The resulting graph is sparse but connected and representative nodes are highlighted. Moreover, the proposed method does not depend on density parameters estimation. Therefore, the proposed adaptive graph construction method itself has already made a contribution to graph-based learning in general.

In the next study, we present heuristic criteria to select representative data instances for labeling. Representatives are studied in diverse network problem such as resilience and epidemiology (Albert, Jeong, & Barabsi, 2000; Holme, Huss, & Jeong, 2002; Pastor-Satorras & Vespignani, 2002). Scale-free like networks, for example, are known to be resilient to random attacks but very sensitive to attacks targeting on hubs. Hubs may be considered representatives in this case (Albert et al., 2000; Cohen, Erez, ben Avraham, & Havlin, 2000; 2001; Jeong et al., 2001; Newman et al., 2002). Other results stand out high betweenness nodes, for example (Holme et al., 2002). That is, centrality measures are able to characterize representatives in this realm. However, there still lacks study of such an important issue in machine learning. Especially, in semi-supervised learning, we have a task to label some data instances. The question is: which data instances should be labels? Previous studies only concern the question of "quantity" or "percentage" of data instances which should be labeled and all data instances are considered to have same importance. In this paper, we go further to investigate the data heterogeneity in semi-supervised learning. Some preliminary results have been obtained in Refs. Araújo and Zhao (2013a, 2013b). Therein, the authors analyze networked data model and propose representative nodes selection strategies using various graph centrality measures. The hypothesis is that representative nodes may correspond to high (or low) score of some network measures. In that study, two network models are considered: Girvan–Newman's random clustered network (GN) and Lancichinetti–Fortunato–Radicchi's clustered networks (LFR) (Danon, Daz-Guilera, Duch, & Arenas, 2005; Girvan & Newman, 2002; Lancichinetti, Fortunato, & Radicchi, 2008). The authors found that the score of clustering coefficient measure of each node stands out for GN networks and the betweenness as well as degree related measures stand out for LFR networks. The authors conclude that these criteria may be useful to real world networks presenting the features of GN or LFR networks. Besides that, the authors proposed the aggregation of centrality measurements through principal components analysis as a unified criterion, among which the second principal component stands out. In this paper, we continue to study this topic. Specifically, we study more network measures, such as Katz index, random walk closeness (RWC), harmonic closeness (Ch), and hierarchical betweenness variation (Hbtw). Moreover, we characterize the relationship between representative nodes and the scores of network measures in hierarchical structure of networks. Extensive computer simulations are performed considering various datasets, various network construction methods, and applying to general network models instead of only GN and LFR. Our hypothesis is that the graph structure may highlight representative nodes and their selection for labeling by hand may improve SSL performance as well as minimize the manual labeling cost (Araújo & Zhao, 2013a; 2013b). Although the importance of this research topic is quite intuitive, we have not found works in the literature to explicitly treat this problem yet. Therefore, our study provides a way toward the understanding of heterogeneous structure of datasets in machine learning.

The rest of this paper is organized as follows. Section 2 presents the problem statement, related works and discussion on classic and state-of-the-art graph construction methods. Section 3 describes the proposed method and some of its benefits including reasonable computational cost. Section 4 discusses representatives characterization through network centrality measures and unified approaches to. Benchmarks, methods, parameters settings and simulations are described in Section 5. Proposed method is firstly evaluated in Section 6. And results for SSL seeded by representatives are presented and discussed in Section 7. Finally, Section 8 presents concluding remarks and points out some open issues and future works.

2. Formulation and related works

Let us consider an unlabeled vector-based dataset, denoted by $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \subset \mathbb{R}^D$. The semi-supervised learning consists of the following steps:

Step 1. A few data instances, $\mathbf{X}_l = \{\mathbf{x}_1, \ldots, \mathbf{x}_l\} \subset \mathbf{X}$, are arbitrarily chosen by a human expert in the field under study with a constrained budget, l;

Step 2. For each data instance $\mathbf{x}_i \in \mathbf{X}_l$, the expert provides a label y_i according to the domain of the application, $y_i \in \{1, \ldots, \kappa\}$, where κ is the number of classes;