Information Sciences 460–461 (2018) 293–317

Contents lists available at ScienceDirect

Information Sciences

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

Active learning via collective inference in network regression problems

Annalisa Appice[a,b,c,*], Corrado Loglisci[a,b], Donato Malerba[a,b,c]

[a] Department of Informatics, Università degli Studi di Bari Aldo Moro, via Orabona, Bari 4 - 70125, Italy
[b] Consorzio Interuniversitario Nazionale per l'Informatica - CINI, Italy
[c] Centro Interdipartimentale di Logica e Applicazioni - CILA, Italy

ARTICLE INFO

Article history:
Received 3 February 2017
Revised 23 January 2018
Accepted 12 May 2018

Keywords:
Network regression
Active learning
Collective inference
Correlation analysis

ABSTRACT

Active learning is a promising machine learning paradigm for querying oracles and obtaining actual labels for particular examples. Its goal is to decrease the number of labels needed, in order to learn a predictive model able to achieve a high level of accuracy. It may turn out to be advantageous in several regression problems where scarce labels can be acquired. A novel active learning algorithm for regression problems in network data is defined. This algorithm performs active learning by taking into account explicitly the correlation property of network data, which makes the labels of linked nodes related to each other. Specifically it resorts to collective inference, in order to accommodate the data correlation in the active selection of the network nodes labeled by oracles. The empirical study proves that the proposed combination of active learning and collective inference can actually boost regression performances in various network domains.

1. Introduction

Artificial intelligence (AI) has been considerably developed during the last thirty years. Thus far, a wide number of artificial intelligence systems has provided various machine learning algorithms for accurate predictive modeling (e.g. [2,3,22,33]). They commonly find a model in data and predict unseen values using that model. As computer systems continue to become more and more powerful and complex by collecting huge volumes of data - often with a complex structure - the capabilities of artificial intelligences will also increase. Network data (e.g. sensor networks, communication and financial networks, web and social networks) are becoming an increasingly important challenge in AI [37]. Regardless of where we encounter them in day-to-day life, network data consist of nodes, which may be connected to each other by edges. The nodes of a network arising from a peculiar domain are, generally, of the same type, i.e. they are described by a vector of fixed properties. The nodes of networks arising from different domains, however, may be associated with different properties. For example, the nodes of a social network (e.g. Twitter) are associated with social properties (e.g. number of tweets, number of followers), while the nodes of a spatial network (e.g. a solar photovoltaic grid) are associated with geospatial properties (e.g. solar radiation, temperature). The edges between the nodes may express an explicit relation, which reflects the dependence between the properties of the nodes. Friendship is an example of relation in social networks, while geographical closeness is an example of relation in spatial networks. This study proposes a machine learning algorithm for predictive modeling in a

[*] Corresponding author at: Department of Informatics, Università degli Studi di Bari Aldo Moro, via Orabona, Bari 4 - 70125, Italy.
E-mail addresses: annalisa.appice@uniba.it, appice@di.uniba.it (A. Appice), corrado.loglisci@uniba.it (C. Loglisci), donato.malerba@uniba.it (D. Malerba).

https://doi.org/10.1016/j.ins.2018.05.028
0020-0255/© 2018 Elsevier Inc. All rights reserved.

data network. The model found in a training data network can be used to predict unseen data of a new testing network that arises from the same training domain (i.e. both training and testing nodes are described by a vector with the same properties; they are linked according to the same type of relation).

Predictive modeling of network data is made complex due to the presence of *correlation*. This is a deterministic or probabilistic dependence between the values of a property on linked nodes [18]. It is apparent in the positive form in a wide variety of network domains like social and spatial domains [29,49]. In social data analysis, correlation can be recognized in the homophily principle, that is, the tendency of nodes with similar values to be linked with each other [32]. In spatial data analysis, correlation can be recognized in Tobler's first law of geography, that is, the tendency of a geophysical attribute to take a value at a given location that may be similar to the values of that attribute in nearby locations [27]. Recent studies [6,12,30,36,45,49] have shown that taking label correlations into account may contribute to improving the accuracy of predictive inferences in network data domains [36,49]. In this context, *collective inference* algorithms, that reason collectively by predicting labels of linked examples simultaneously, offer a unique opportunity to accommodate label correlations in the learned models [36]. Although most work on collective inference is defined for classification problems, a few collective inference algorithms have recently been proposed for regression problems [6,29]. In any case, these studies do not pay any attention to the procedure to acquire a representative labeled set from the data network, while this is often the first step towards performing accurate predictive inference, even when few labels are acquired [20].

The widely-used paradigm for label collection is called passive learning, where training samples are randomly selected from the underlying distribution and manually annotated by an oracle (e.g. human experts). However, due to the high cost associated with the above label collection process, it often happens that there are not enough labeled samples to train a high quality predictive model. An important research question is to develop algorithms that learn an accurate model with minimal labeling effort required in such tasks. One promising learning strategy is to use *active learning*. In this strategy, rather than being presented with a labeled training set from the start, the learner is allowed to request labels for particular examples, with the goal of decreasing the number of labels needed to achieve the desired level of accuracy. At present, various active learning algorithms have been investigated for network classification [7,26,31,53]. Few active learning algorithms have already been developed for regression [9–11,16,39]. A seminal study [25] combines active learning and correlation analysis of univariate linked numeric data. However, to the best of our knowledge, active learners that take direct advantage of correlation in linked multivariate data have still not been considered for network regression problems.

The main contribution of this study is the description of a novel holistic strategy, where collective inference is used to drive the active learning for network regression. A novel algorithm, called CoNeRa (Collective Network Regression via Active learning) is described. It performs predictive modeling for a numeric target (also called label) in an initially unlabeled training data network. Specifically it iteratively selects a budget of training nodes to be labeled by the oracle, so that a final accurate regression model can be learned from this partially oracle-labeled network. This model can be used to predict unseen data in a testing data network that is acquired in the same domain condition of training.[1] This algorithm uses both the descriptive information (node properties) and the network structure (correlation property) during the training procedure. According to the collective inference theory, the algorithm learns a collective regression model by accounting for descriptive data associated with nodes, as well as collective data yielded by handling the property of label correlation throughout the network. According to the active learning theory, the algorithm requests labels for particular examples, which are selected based upon a disagreement measure. Specifically, the disagreement quantifies the correlation of the (potential) labels surrounding a (potential) target value. It is noteworthy that this algorithm is based on regression, as the final goal is performing predictive modeling of a numeric target. However, it implements an active learning procedure that also uses clustering, that is a form of unsupervised classification. Specifically, the active learning component integrates a new constraint-based clustering solution. This component is able to discover a cluster structure that depicts the correlation observed throughout the network in the descriptive data (i.e. each cluster represents a region of connected nodes associated with similar descriptive data). Assuming a dependence between the descriptive variables and the target variable, this cluster knowledge, discovered as a model of the descriptive data correlation, should implicitly provide some information on the correlation property of the target. Under this assumption, the active learning procedure considers this cluster knowledge, in order to guarantee diversity in the label acquisition and avoid over-investing in descriptive areas of the training data, which have already been explored. Finally, the collective regression model, induced from the final set of examples labeled by the active learning, can be used in a testing procedure involving unseen nodes.

The paper is organized as follows. Section 2 summarizes the main research contribution of this study, while Section 3 reports relevant related work. Section 4 illustrates the proposed collective active learning algorithm and its time complexity. Section 5 describes the datasets, the experimental methodology and reports the results. Finally, in Section 6 some conclusions are drawn and future work is outlined.

2. Research contribution

The idea of combining collective inference with active learning is not new in machine learning research; indeed, it has already been used for the classification of network data [7]. Theoretically, the collective active learning strategy defined

[1] The presented algorithm is formulated, in order to process data acquired at a specific time without accounting for mechanisms of incremental modeling of historical data. This is now out of the scope of this paper, although it may represent an interesting future direction of investigations.