Information Sciences 425 (2018) 107–126

Contents lists available at ScienceDirect

Information Sciences

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

Multi-type clustering and classification from heterogeneous networks

Gianvito Pio[a,*], Francesco Serafino[a], Donato Malerba[a,b], Michelangelo Ceci[a,b]

[a] Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
[b] Big Data Laboratory, National Interuniversity Consortium for Informatics (CINI), Rome, Italy

ARTICLE INFO

Article history:
Received 10 January 2017
Revised 3 October 2017
Accepted 7 October 2017
Available online 9 October 2017

Keywords:
Heterogeneous networks
Multi-type clustering
Multi-type classification

ABSTRACT

Heterogeneous information networks consist of different types of objects and links. They can be found in several social, economic and scientific fields, ranging from the Internet to social sciences, including biology, epidemiology, geography, finance and many others. In the literature, several clustering and classification algorithms have been proposed which work on network data, but they are usually tailored for homogeneous networks, they make strong assumptions on the network structure (e.g. bi-typed networks or star-structured networks), or they assume that data are independently and identically distributed (i.i.d.). However, in real-world networks, objects can be of multiple types and several kinds of relationship can be identified among them. Moreover, objects and links in the network can be organized in an arbitrary structure where connected objects share some characteristics. This violates the i.i.d. assumption and possibly introduces autocorrelation. To overcome the limitations of existing works, in this paper we propose the algorithm HENPC, which is able to work on heterogeneous networks with an arbitrary structure. In particular, it extracts possibly overlapping and hierarchically-organized heterogeneous clusters and exploits them for predictive purposes. The different levels of the hierarchy which are discovered in the clustering step give us the opportunity to choose either more globally-based or more locally-based predictions, as well as to take into account autocorrelation phenomena at different levels of granularity. Experiments on real data show that HENPC is able to significantly outperform competitor approaches, both in terms of clustering quality and in terms of classification accuracy.

1. Introduction

Many objects and data in the real world can be considered interconnected (i.e., through relationships, interactions, etc.), forming complex information networks. Information networks can be found in several social, economic and scientific fields, ranging from the Internet to social sciences, including biology, epidemiology, geography, finance and many others. Current studies about mining networked data mainly focus on homogeneous information networks [28,31,41], i.e., networks composed of a single type of object and a single type of link. However, in real life scenarios, there could be multiple types of object, connected to each other through different kinds of link, forming heterogeneous information networks. Examples

* Corresponding author.
E-mail addresses: gianvito.pio@uniba.it (G. Pio), francesco.serafino@uniba.it (F. Serafino), donato.malerba@uniba.it (D. Malerba), michelangelo.ceci@uniba.it (M. Ceci).

https://doi.org/10.1016/j.ins.2017.10.021
0020-0255/

can be found in biology, where genes, proteins, organisms, etc. are associated to each other through different types of link. Another example is in the analysis of bibliographic data, where different types of link exist among authors, conferences, journals and papers.

These considerations motivate the recent interest in mining heterogeneous information networks, which in most cases focus on the clustering task. Typically, (*multi-type*) clustering is based on both the attribute values of the objects (possibly of different types) and the links among them (e.g., spectral clustering [26], LinkClus [49] and CrossClus [50]). Clustering has been also exploited as a preliminary phase for other data mining tasks, such as link prediction [3] and semantic tagging [46], as well as for the construction of higher-level features or multiple views of the data [8]. Further interesting examples can be found in studies that proposed ranking-based clustering approaches (e.g., RankClus [44] and NetClus [45]), that generate efficiently results for both ranking and clustering. Some other methods perform classification [18,19]. They take advantage of links in heterogeneous information networks to propagate knowledge across the nodes, enforcing the similarity between similar objects, if linked. Approaches that work in this direction are based on label propagation [51] and collective classification [39].

One of the main problems that these approaches tackle when learning predictive models from network data (regardless of the type of network, that is, homogeneous or heterogeneous) is that data are affected by some form of autocorrelation [1,41]. This means that the value of an attribute at a given node depends on the values of the same attribute of the nodes it is connected with. This phenomenon is a direct violation of the assumption that data are independently and identically distributed, which is at the basis of most data mining methods. At the same time, autocorrelation also offers a unique opportunity to improve the performance of predictive models on network data, since inferences about one object can be used to improve inferences about related objects. Autocorrelation can be recognized in several fields, for instance, in spatial data analysis it can be recognized in the (Tobler's) first law of geography: "Everything is related to everything else, but near things are more related than distant things". In social analysis, autocorrelation can be recognized in the homophily principle [29], which shows that people connected through friendship relationships tend to share many sociodemographic, behavioral, and intra-personal characteristics. This is also important in marketing [9].

In this paper, we propose a method which is able to perform both clustering and classification tasks on heterogeneous networks. In particular, it is able to group together heterogeneous objects in a network and to assign labels to unlabeled objects, implicitly taking into account autocorrelation in a collective learning setting. Similarly to multi-type clustering, where cluster labels are associated to objects of multiple types in an unsupervised fashion, in our work we simultaneously cluster and classify objects of different types. Classification can be performed according to different classification schemes i.e. it can be single-label, multi-label, hierarchical, hierarchical multi-label (HMC), etc. In this work we focus attention on the single-label setting and we call the considered learning task *multi-type clustering and classification* from heterogeneous networks. This task is not completely new in the literature and has connections with the task of multiple predicate learning [36] in ILP. The difference is that it is applied to heterogeneous networks and not to logic clauses. Connections can also be found with the task of multi-label collective classification [23,38], where, however, objects to be classified are of the same type.

In this work, we consider the *within-network* setting [11]: objects for which the class is known are linked to objects for which the class must be estimated [28]. This setting is semi-supervised and differs from the *across-network* setting, where learning is performed from one (labeled) network and prediction is performed on a separate, presumably similar network [27,41].

In order to simultaneously consider the clustering and the classification tasks, the solution we propose is based on predictive clustering [5], which combines elements from both tasks and allows us to properly take into account the autocorrelation phenomenon: Clusters of similar objects are identified, and a cluster description and a predictive model are associated to each cluster. Unlabeled objects are assigned to clusters on the basis of the cluster descriptions and the corresponding predictive models are considered to provide predictions for the target property. The basic idea is to build (possibly overlapping) clusters of heterogeneous nodes of the network, such that autocorrelation can be implicitly considered when learning the classifier. This means that the clustering algorithm should preserve the network structure, that is, linked objects should have similar cluster membership [43]. In other words, highly connected objects should fall in the same clusters, contributing to learning the same predictive models. When exploiting multi-type clustering for (multi-type) classification, we use the intuition that class values of objects of type A are in some way related to class values of objects of type B, when all these objects belong to the same heterogeneous cluster (or sub-network). See Fig. 1 for an example.

The way the clusters are generated implicitly influences prediction. In fact, a hierarchical organization of clusters, which is learned in this work, facilitates, from the descriptive perspective, the understanding of the results by human experts. From the predictive perspective, this organization results in the possibility of choosing either more globally-based or more locally-based predictions. This is because each cluster can naturally consider different effects of the autocorrelation phenomena on different portions of the network: at higher levels of the hierarchy, clusters will be able to consider autocorrelation phenomena that are spread all over the network, while at lower levels of the hierarchy, clusters will consider the local effects of autocorrelation.

Another characteristic that can influence the prediction is the extraction of overlapping clusters. Indeed, overlapping clusters give the opportunity to base predictions on more than one heterogeneous sub-network. For this reason, we extract hierarchically organized and possibly overlapping heterogeneous clusters.

Since clusters are used for predictive purposes, a clustering algorithm which takes as input the number K of clusters would affect the predictive capabilities of the learned models. In particular, setting K equal to the number of classes would