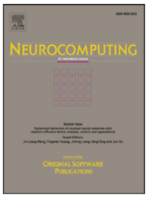
Contents lists available at ScienceDirect

Neurocomputing

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



MTDE: Multi-typed data embedding in heterogeneous networks

Haifeng Sun*, Jianxin Liao, Jingyu Wang, Qi Qi

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, PR China

ARTICLE INFO

Article history:
Received 18 January 2017
Revised 14 May 2017
Accepted 20 June 2017
Available online xxx

Keywords:
Non-Gaussian distribution
Unsupervised learning
Embedding
Heterogeneous networks
Gibbs sampling

ABSTRACT

Vectorized representations as an important data representation way play an essential role in many data mining applications. Now, more and more applications are based on multi-typed information network, such as social networks, which is called heterogeneous networks. However, most data in heterogeneous networks are far from Gaussian distribution. Gaussian models are inappropriate choices to model such data. On the other hand, most traditional embedding methods are based on single typed data, and cannot be directly applied in data with network structures. In this paper, we propose an embedding method, named as Multi-typed Data Embedding (MTDE), vectorized represents the data in non-Gaussian distribution. It achieves Latent Spaces for every typed data and a multi-typed latent translational space by a probabilistic model based on Gibbs sampling method. First, it embeds the objects in network not only considering the relationships in same typed data, but also the network structure. Second, it provides a translational space to make the comparison of different typed data available. Thus, we can utilize MTDE to compare different typed data in more data mining applications. Our experiments on DBLP show that MTDE learns high-quality embedding. Moreover, other data mining tasks, e.g. Clustering, based on MTDE achieve a better performance than the state-of-the-art methods.

1. Introduction

Data representations frequently arise in non-Gaussian distributed data. They are easier to handle since each data can be viewed as a point residing in an Euclidean space [1–3]. Similarities between different data points can be directly measured by an appropriate matrix to solve traditional tasks such as classification, cluster and retrieval [4]. For example, word2vec [1], which provides a vectorized representation method for words, makes deep neural network learning methods can be more easily used in the textual information [5,6]. As shown in [7], learning good representations is one of the fundamental problems in data mining. It often has a stronger impact on performance than designing a more sophisticated model.

Some studies model textual and image data based on Dirichlet distributions, such as the nonlinear transformation based methods [8], use decorrelation strategy to vectorized data representations in a neutral latent space. Unfortunately, many networked data source (e.g. DBLP, Twitter) are multi-typed data combined in graphs. This kind of methods are hard to handle multi-typed data [9]. Current research has focused on either pre-defined feature vectors [10] or sophisticated graph-based algorithms [11]. The development of a vectorized representation for networked data in is of great importance to encode content. Once the vectorized representation is obtained, the network mining tasks can be readily solved by the machine learning algorithm [9].

The Multi-typed data network is different from single-typed data network. First, most data in information network, such as social network, does not exist in isolation, but in combination with various data types [12]. These interactions can be formed either explicitly or implicitly with the linkages between them. For example, in a bibliographic information network, the information about authors of academic articles is highly related to the topics of academic research papers. And a large number of authors more probability published their papers in a specific set of conferences or journals. There are three types information, *authors, papers* and *publishers* in bibliographic information network. They construct a heterogeneous network [13]. Second, good vectorized representations for multi-typed data should not only represent the same typed data relationship but also the different typed data relationship. The similarity of representations between an author and his published work should much closer than the work of other authors in a bibliographic network.

To address the aforementioned challenges, we present a novel approach on network representation learning, termed *Multi-typed Data Embedding* (MTDE), which jointly considers both the objects as well as the relationship among them. It maps the same typed

* Corresponding author.
E-mail addresses: hfsun@bupt.edu.cn (H. Sun), jxlbupt@gmail.com (J. Liao), wangjingyu@bupt.edu.cn (J. Wang), qiqi8266@bupt.edu.cn (Q. Qi).

http://dx.doi.org/10.1016/j.neucom.2017.06.079
0925-2312/© 2017 Elsevier B.V. All rights reserved.

objects into a latent space so that the objects can be represented into a same space where they can be directly compared. Objects of different types can be mapped in different spaces where they can be compared by space transforming. Different from traditional linear embedding models, such as HOSVD [14], proposed method uses a generative probabilistic modeling, data arises from a generative process that includes latent factors. The generative process defines a joint probability distribution over both the observed and latent random factors [15]. We perform data analysis by using that joint distribution to compute the conditional distribution (or posterior distribution) of the latent factors given the observed data. In MTDE, the observed data are the objects in heterogeneous networks, such as the words in a paper and the authors in a paper; the latent factors are the hidden variables, such as the topics for words and the latent interests for authors. The computational problem of inferring the hidden latent space structures from the heterogeneous networks is the problem of computing the posterior distribution, the conditional distribution of the latent factors given the network.

Generally, the contribution of this paper is highlighted as follows:

1. Proposed an unsupervised probabilistic model jointly learning latent space structures of multi-typed data. The unsupervised probabilistic model learns the low-dimensional latent spaces for each typed objects. The mapping of objects to latent spaces can be used as the data representation. It improves the performance of data representation by leveraging the multi-typed interaction to learn the latent space. And it is suitable for many network orientated data mining applications.

2. Proposed a method to compare different typed data. In most previous work, the different typed data cannot compare directly. However, the relationship among different typed objects can be used to calculate the semantic distances in heterogonous network. A tensor in MTDE obtains the relationship of different typed latent spaces, we can use the mapping among latent spaces to compare the different typed objects.

2. Related work

A branch of latent feature embeddings is motivated by applications such as collaborative filtering and link prediction in networks that model the relations between entities from latent attributes [16]. These models often transfer the problem as learning an embedding of the entities, which corresponds to a matrix factorization problem of observed relationships. Some researchers proposed a joint factorization approach on both the linkage adjacency matrix and document-term frequency for Web page categorizations [17,18]. In addition, Shaw et al. [19] proposed a structure preserving embedding framework that embeds graphs to a low-dimensional Euclidean space with global topological properties preserved. DeepWalk [20] learned latent representations of vertices in a network from truncated random walks. However, these models focus only on single relations that do not adapt to heterogeneous settings. A natural extension of these methods to heterogeneous settings is by stacking multiple relational matrixes together, and then applying a conventional tensor factorization [21–23]. The disadvantage of such multi-relational embeddings is the inherent sharing of parameters between different terms, which does not scale to large graphs.

Since the topic modeling proposed by Blei et al. [24], the Beta distribution models [25] and Dirichlet Distribution models [26] are widely studied in textual content information. In [27], the authors provided a noval variational inference method to inference the latent space from observed data. Some textual content heterogeneous networks integrate the topic modeling with the network structures to embedding the network based on the topics of textual content. For example, the author-topic model [28] studied both the document embedding and the author embedding based topic distribution of content. VideoTopic [29] utilized the content description, such as metadata, to capture user interests in video by using a topic model to represent the video, and then generated recommendations by finding those video that most fit to the topic distribution of user interests. Topic models are also applied on other typed data, such as Youngchil et al. [30] used topic models to study the relationship graph of popular social network. They applied the topic model to cluster the purely on the social network graph consisting of following edges.

3. Multi-typed data embedding

In this section, we introduce the related concepts and define the problem of MTDE.

Definition (Heterogeneous Information Network). Given a set of objects from N types $X = \{X_n\}_{n=1}^{N}$, where X_n is a set of objects belonging to n_{th} type, a weighted graph $G =< V, E, W >$ is called an information network on objects X, if $V = X$, E is a binary relation on V, and $W : E \rightarrow R^{+}$ is a weight mapping from an edge $e \in E$ to a real number $w \rightarrow R^{+}$. Specially, we call such an information network heterogeneous networks when $N \geq 2$; and homogeneous network when $N = 1$.

Many information networks in real applications could be described by heterogeneous networks. For example:

Example 1 (Bibliographic information network)**.** A bibliographic network consists of rich information about academic papers, each using a set of contents, written by a group of authors, published in a venue (a conference or a journal). Such a bibliographic network is composed of three types of objects: author, venues, and textual content. The topological structure of a paper information network forms a heterogeneous network. We can use the data of network to construct a three-order tensor G, a value in this tensor represent the number of times the corresponding author write the corresponding words published in the corresponding venue.

In the topic modeling perspective, document content is based upon the idea that the probability distribution over words in a document can be expressed as a mixture of topics, where each topic is a probability distribution over words [24]. In Fig. 1, the matrix A represents topics probability distribution of a set of documents. Each column of matrix A is a topics probability distribution of corresponding paper. The matrix T is a topic-word matrix represents the topic distribution of each word. Each column of T represent a topic probability distribution of corresponding word. The matrix G is a word-document frequency matrix, which represent the distribution of each word in a document. The topic modelings can be represented as:

$$A = TG \tag{1}$$

Thus, the topics probability distribution is an embedding method to encode the words and documents. By the same way, the probability distribution over authors of a paper can be represented as a mixture of latent space of authors interests, where each latent space of interest is a probability distribution over authors; the probability distribution over venue of a document can

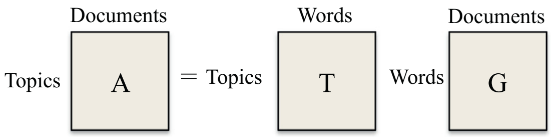


Fig. 1. Topic model in a matrix decomposition perspective.