Pattern Recognition Letters 000 (2016) 1–8

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

The mutual information between graphs☆

Francisco Escolano[a,*], Edwin R. Hancock[b], Miguel A. Lozano[a], Manuel Curado[a]

[a] Department of Computer Science and AI, University of Alicante 03690, Spain
[b] Department of Computer Science, University of York, YO10 5GH United Kingdom

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Graph entropy
Mutual information
Manifold alignment

ABSTRACT

The estimation of mutual information between graphs has been an elusive problem until the formulation of graph matching in terms of manifold alignment. Then, graphs are mapped to multi-dimensional sets of points through structure preserving embeddings. Point-wise alignment algorithms can be exploited in this context to re-cast graph matching in terms of point matching. Methods based on bypass entropy estimation must be deployed to render the estimation of mutual information computationally tractable. In this paper the novel contribution is to show how manifold alignment can be combined with copula-based entropy estimators to efficiently estimate the mutual information between graphs. We compare the empirical copula with an Archimedean copula (the independent one) in terms of retrieval/recall after graph comparison. Our experiments show that mutual information built in both choices improves significantly state-of-the art divergences.

1. Introduction

1.1. Motivation

One of the key elements for building a pattern theory is the definition of a set of principled dissimilarity measures between the mathematical structures underpinning the theory. For instance, in vectorial pattern recognition, one of the fundamental degrees of freedom of an information theoretic algorithm (for clustering, matching, classification and learning) is the choice of a divergence. There are some possibilities including mutual information, Kullback–Leibler, Bregman divergences, and so on (see [11] for a review).

The mutual information $I(X; Y)$ between two variables X and Y is very interesting since it captures high-order statistical dependencies between the variables. However, when these variables are graphs we must address two issues. Firstly, we must express graphs X and Y as random variables, beyond the simplistic model of Erdös-Rényi model. In such model a *random graph* is built by assigning a probability to the edges. However this model does not fully characterize the probability that a given graph (with a variable number of vertices) is observed. Secondly, since $I(X;Y) = H(X) + H(Y) - H(X,Y)$ we must estimate the Shannon entropy $H(.)$ of a graph. There are several approaches for estimating graph entropy. The most efficient entropy estimators rely on functionals aiming to quantify the amount of information flowing through the graph. For instance, in [1] the state vector of the steady state random walk on the graph defines a discrete probability function on the nodes. The Shannon entropy of such a probability function yields $H(.)$. On the other hand, quantum walks probing is used in [31] for providing mixed quantum states known as density matrices. Following Passerini and Severini [26], the von Neumann entropy (or quantum entropy) maps discrete (graph) Laplacians to quantum states: scaling the graph Laplacian by the inverse of the volume of the graph we obtain a density matrix whose entropy can be computed using the spectrum of the discrete Laplacian. More recently Han et al. [14] have approximated the von Neumann entropy by formulating it in terms of node degrees.

The above methods for estimating graph entropy operate on the graph itself, i.e., they consider the graph as a coder of node/vertex dependencies and describe entropy in terms of its capability for diffusing information. However, in this paper we consider that a graph is a special type of random variable with a bounded number of nodes and/or edges and we model structural distortion in terms of a novel coding (transforming graphs into low-dimensional manifolds). Then, it is possible to exploit the apparatus of bypass entropy estimators [17,23]. In fact, bypass estimators do not rely on estimating probability density functions but on Euclidean distances between vectorial patterns. This means that the Parzen approximation of the probability density function is no longer needed since entropy can be estimated directly from the samples.

☆ This paper has been recommended for acceptance by Cheng-Lin Liu.

* Corresponding author. Fax: +34 965903902.
E-mail address: sco@dccia.ua.es (F. Escolano).

http://dx.doi.org/10.1016/j.patrec.2016.07.012
0167-8655/

On the other hand, the development of graph embeddings which map vertices to multi-dimensional spaces *bypasses the rigid discrete representation of graphs*. After being embedded, the associated multi-dimensional subspace must retain the rich topological information of the original representation. Many embeddings have been proposed so far: ISOMAP [30], Heat Kernels [33] , Diffusion Maps [16], Laplacian Eigenmaps [2], Commute Times [27], Centered Normalized Laplacian [28] among others. Most of the these latter structure preserving embeddings (i.e. distances in the embedding are correlated with structural properties) establish a formal link between topology (usually encoded in spectral terms) and some kind of metric or dissimilarity measure in the subspace. Understanding and exploiting the latter formal link is key to quantifying the effectiveness of the corresponding embedding for a given task. For instance, graph comparison. In [9] there are experimental graph comparisons showing that the Commute Time (CT) embedding outperforms the alternatives in terms of retrieval/recall for the best dissimilarity measure in a given set. In addition, the fact that the latter embedding induces a metric allows us to work in the multi-dimensional subspace of the embedding. Here, problems such as finding graph prototypes are more tractable. It is then possible to return to the original topological space via inverse embedding [8].

1.2. Contribution

With these ingredients at hand (bypass estimators and suitable embeddings), the mutual information between graphs can be defined in terms of *structural information channels* (Section 2). In such channel model, there will be embedding-based encoders and inverse embedding decoders. The channel will be characterized by a conditional entropy relying on a global non-rigid transformation between the input embedding and the distorted one. We will devote Section 3 to present how to obtain a multi-dimensional estimation of Mutual Information (MI) from the combination of copulas and Rényi entropy estimators. In Section 4 we will compare MI for embedded graphs with other challenging dissimilarities. In order to perform a fair comparison we will use the GatorBait database which has been proven to be a very challenging one despite its small size. This is due to the fact that it exhibits we very high intra-class variability and very low inter-class variability in only 100 samples. In Section 5 we will present our conclusions and future work.

Our main contribution in this paper is to define graph similarity through a model of structural information channel where distortion relies on manifold and MI is estimated through different types of copula functions.

2. Information channels and manifold deformation

Let $\mathcal{X} = (\mathcal{V}_X, \mathcal{E}_X)$ be a random variable $\mathcal{X} : \Omega \to E$ defined over the set of unweighted and undirected graphs Ω with node-sets $\mathcal{V}_X$ having $|\mathcal{V}_X| = n$ nodes. Then, its associated edge-set $\mathcal{E}_X \subseteq \mathcal{V}_X \times \mathcal{V}_X$ satisfies $|\mathcal{E}_X| \leq \binom{n}{2}$ and a realization of $\mathcal{X}$ is given by an $n \times n$ adjacency matrix $\mathcal{A}_X \in E$.

Let $K_X : \mathcal{V}_X \times \mathcal{V}_X \to \mathbb{R}$ be a topological similarity measure $K_X(i, j)$, ideally a kernel, between two nodes $i, j \in \mathcal{V}_X$. We assume that the probability mass $p(\mathcal{X})$ relies on the probability mass of $K_X(.,.)$ as follows: peaked similarity distributions yield less probable realizations than flat ones. This choice is convenient for two reasons. Firstly, it is consistent with recent definitions of graph entropy (see [10,26] and [14]). Secondly, it provides a principled framework for understanding graph distortion in terms of the distortions induced in $K_X(.,.)$.

Let $\mathcal{C}$ be an *structural information channel* $\mathcal{X} \to \mathcal{C} \to \mathcal{Y}$ where $\mathcal{Y} = (\mathcal{V}_Y, \mathcal{E}_Y)$ satisfies $\mathcal{V}_Y = \mathcal{V}_X$. Then, the conditional probability $p(\mathcal{Y}|\mathcal{X})$ describes a noiseless channel with respect to the vertices or nodes, but a noisy channel with respect to the edges. The channel $\mathcal{C}$ generates structural noise (insertions and/or deletions of edges) through an unknown matching function $g : \mathcal{E}_X \to \mathcal{E}_Y \bigcup \{\Phi\}$, where Φ is the NULL label accordingly with Myers et al. [20]. Finding the function $g(.)$ is typically posed in terms of minimizing the graph-edit distance between $\mathcal{X}$ and $\mathcal{Y}$ (see [29]). Although many recent developments have proposed approximations of the graph-edit distance (see for instance [12]) they are (to some extent) rooted in marginalizing $p(\mathcal{Y}|\mathcal{X})$. Marginalization tends to capture or preserve local coherence between the matched edges at the cost of losing global coherence, especially when the input graphs $\mathcal{X}$ and $\mathcal{Y}$ are unattributed.

Here, we propose a different approach which enforces global coherence. Let $f_{\mathcal{X}} : \mathcal{V}_X \to \mathbb{R}^d$, with $d \ll n = |\mathcal{V}_X|$, a graph embedding function. The embedding $f_X(.)$ induces a manifold $\mathcal{M}_X$, i.e. a subspace of $\mathbb{R}^d$, where the structural similarities $K_X(i, j)$ between pairs of vertices $i, j \in \mathcal{V}_X$ are encoded by a geodesic. Graph embedding functions are such that the Euclidean norm $||f_{\mathcal{X}}(i) - f_{\mathcal{X}}(j)||^2$ is a reasonable approximation of the geodesic insofar d matches the intrinsic dimension of the manifold (see [9]).

Therefore, since a graph $\mathcal{X}$ is mapped to a subspace/manifold $\mathcal{M}_X \subseteq \mathbb{R}^d$ we assume that the embedding function $f_X(.)$ plays the role of an encoder associated with the channel $\mathcal{C}$ which transmits one manifold $\mathcal{M}_X$ at a time. Given a manifold to transmit, its encoding is not free of error, i.e. it is noisy: different vertices can be mapped to the same point of $\mathbb{R}^d$. However, we assume that the messages (resulting from the encoding) retain the global topology of their respective graphs $\mathcal{X}$. A simple model for the the conditional distribution $p(\mathcal{Y}|\mathcal{X})$ governing $\mathcal{C}$ is the usual factorization

$$p(\mathcal{Y}|\mathcal{X}) = \prod_{i=1}^{n} p\left(\Theta_Y^{(i)}|\Theta_X^{(i)}\right), \tag{1}$$

where $\Theta_Y^{(i)}$ and $\Theta_X^{(i)}$ are respectively the $i-$th points of manifolds $\mathcal{M}_Y$ and $\mathcal{M}_X$. However, the above factorization is misleading, since we have

$$p(\Theta_Y^{(i)}|\Theta_X^{(i)}) \propto \exp\left\{-\frac{1}{2}\left|\left|\frac{\Theta_X^{(i)} - \mathcal{T}(\Theta_Y^{(i)}; \mathbf{W})}{\sigma}\right|\right|^2\right\}, \tag{2}$$

where $\mathcal{T}(.; \mathbf{W})$ is a *global* non-rigid transformation parameterized by $\mathbf{W}$, and σ is the bandwidth (see [9] for more details). This is consistent with assuming that we cannot observe the matching function $g(.)$ but instead its effects in the similarity matrix $K_X(.,.)$ in order to produce a new one $K_Y(.,.)$ which determines the embedding $f_{\mathcal{Y}} : \mathcal{V}_Y \to \mathbb{R}^d$ leading to $\mathcal{M}_Y$.

The framework developed in this paper encompasses our early research. A model for the information channel $\mathcal{C}$ does not only assume that an output manifold $\mathcal{M}_Y$ is received. It must also specify how it is decoded. We do that through an *inverse embedding*. In our previous work (see [8]) we showed that for certain types of embeddings, e.g. commute-time embeddings, it is possible to approximate $\mathcal{Y}$ with minimal error.

Consequently, in our model we naturally associate distortion (when the information rate exceeds the channel capacity) with excessive deformation, since the capacity of the channel, defined as $C = \max_{p(\mathcal{X})} I(\mathcal{X}; \mathcal{Y})$, decays significantly with the increase of $\epsilon = \sum_{i=1}^{n} ||\Theta_X^{(i)} - \mathcal{T}(\Theta_Y^{(i)}; \mathbf{W})||^2$. This means that, although $\mathcal{T}(.; \mathbf{W})$ is chosen so that ϵ is minimized, the deformation is constrained by a regularization constant, i.e. the channel capacity is bounded by regularization.

Bridging deformation with mutual information $I(\mathcal{X}; \mathcal{Y})$ opens up a way of analyzing structural pattern distortion in terms of rate-distortion theory. In the following section, we propose a means of estimating $I(\mathcal{X}; \mathcal{Y})$ within this framework.

Please cite this article as: F. Escolano et al., The mutual information between graphs, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.07.012