Pattern Recognition 60 (2016) 802–812

Contents lists available at ScienceDirect

Pattern Recognition

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

Statistical graph space analysis

Brijnesh J. Jain

TU Berlin, Fak. IV, Sekr. TEL-14, Ernst-Reuter-Platz 7, 10587 Berlin, Germany

ARTICLE INFO

Article history:
Received 25 November 2015
Received in revised form
18 April 2016
Accepted 25 June 2016
Available online 2 July 2016

Keywords:
Graph edit distance
Graph matching
Fréchet mean
Geometric midpoint
Consistent estimator
Majorize–minimize algorithm

ABSTRACT

The sample mean is one of the most fundamental concepts in statistics. Properties of the sample mean that are well-defined in Euclidean spaces become unclear in graph spaces. This paper proposes conditions under which the following properties are valid: existence, uniqueness, and consistency of means, the midpoint property, necessary conditions of optimality, and convergence results of mean algorithms. The theoretical results address common misconceptions about the graph mean in graph edit distance spaces, serve as a first step towards a statistical analysis of graph spaces, and result in a theoretically well-founded mean algorithm that outperformed six other mean algorithms with respect to solution quality on different graph datasets representing images and molecules.

1. Introduction

Statistical inference deduces properties about a population by analyzing a subset of sampled data. One central path departs from the fundamental concept of mean, then leads via the normal distribution and the Central Limit Theorem to statistical estimation using the maximum likelihood method. The maximum likelihood method in turn is a fundamental approach that provides probabilistic interpretations to many pattern recognition methods.

This central path is well-defined in Euclidean spaces, but becomes unclear in mathematically less structured spaces. Since an increasing amount of non-Euclidean data is being collected and analyzed in ways that have not been realized before, statistics is undergoing an evolution [43]. Examples of this evolution are contributions to statistical analysis of shapes [3,11,27,42], complex objects [48,57], and tree-structured data [14,57].

All these spaces have in common that a well-defined pairwise addition of elements is unknown. The standard approach to generalize the concept of sample mean to arbitrary distance spaces $(\mathcal{X}, d)$ is based on an idea proposed by Fréchet in 1948 [20]. Inspired by Fréchet, a sample mean of n elements $x_1, \ldots, x_n \in \mathcal{X}$ is any element from $\mathcal{X}$ that minimizes the sample Fréchet function

$$F_n(x) = \sum_{i=1}^{n} d(x_i, x).$$

Research on the sample mean of graphs started with the pioneering work by Jiang et al. [40,41]. By using Fréchet's formulation, they studied the sample mean under the term *generalized median* in graph spaces endowed with the graph edit distance [19,45,55].

In principle, any graph distance can be used to study a sample mean of graphs. The majority of research inspired by Jiang et al. [40,41] applied graph distances that can be subsumed as variants and restrictions of the graph edit distance [19,45]. Research predominantly focused on devising heuristics for approximating a sample mean [17,26,31,32,49,51] and on developing central clustering algorithms [4,6,16,24,30,46,47].

In contrast to shape or tree spaces [3,14], there are hardly any studies that aim at understanding elementary theoretical properties of the sample mean in graph spaces. This also includes theoretical issues related to computing a sample mean, which is relevant for applications in statistical pattern recognition, such as in computer vision, network analysis, chemo- and bioinformatics [19,21,45].

The theoretical gap in the field of graph-based representation on the sample mean is prone to misconceptions that hinder statistical inference and resist bridging the gap between structural and statistical pattern recognition. We will point to potential fallacies of the sample mean in graph edit distance spaces in the introductions of Sections 3–5.

In this contribution, we endow graph spaces with a restricted form of graph edit distance, called *graph alignment metric* henceforth. The graph alignment metric is an intrinsic graph metric that occurs in different guises as a widely applied dissimilarity function [7,8,23,56,58]. Then we prove the following properties:

1. Optimization (Section 3):
   (O_1) Existence of sample mean.

E-mail address: brijnesh.jain@gmail.com

http://dx.doi.org/10.1016/j.patcog.2016.06.023
0031-3203/

(O_2) Necessary conditions of optimality for F_n.
(O_3) Convergence results for algorithms that minimize F_n.

2. Statistics (Section 4):
   (s_1) Existence of population mean.
   (s_2) Consistency of sample mean.
   (s_3) Uniqueness of population and sample mean.
3. Pattern recognition (Section 5):
   (p_1) Existence of midpoint.
   (p_2) Coincidence of sample mean and midpoint.

Based on property (O_2) we propose a majorize–minimize mean (MMM) algorithm and establish convergence (O_3) to solutions satisfying necessary conditions of optimality. Section 6 suggests that the theoretically well-founded MMM-algorithm is also useful for applications. In experiments the MMM-algorithm substantially outperformed six other mean algorithms with respect to solution quality on different graph datasets representing images and molecules.

The present treatment is intended as first step along the above mentioned central path of statistical inference in graph spaces. We enter this path by proving the statistical properties (S_1)–(S_3). These properties determine the main purpose of a sample mean as an estimator of a population mean. In addition, properties (S_1)–(S_3) pave the way towards studying the asymptotic behavior of the sample mean as the next step along the central path. First examples of properties at the other end of the central path are the geometric properties (P_1) and (P_2). Both properties induce meaningful and geometrically interpretable update rules for generalizing a subset of pattern recognition methods to graph spaces. To apply these pattern recognition methods, a theoretically sound way to compute a sample mean is desirable. This issue leads to the third set of properties (O_1)–(O_3) related to optimization.

2. Background

This section describes graph alignment spaces as well as the concepts of sample and population mean. Statements in this section are proved in [38].

2.1. Graph alignment spaces

We first introduce attributed graphs and then endow them with an intrinsic metric, called graph alignment metric. We refer to Fig. 1 for explanatory illustrations of some of the concepts we introduce in this section.

Attributed graphs. Let $\mathcal{A} = \mathbb{R}^d$ be the set of node and edge attributes. An attributed graph is a triple $X = (\mathcal{V}, \mathcal{E}, \alpha)$, where $\mathcal{V}$ represents a finite set of nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ a set of edges, and α: $\mathcal{V} \times \mathcal{V} \to \mathcal{A}$ is an attribute function satisfying

1. $\alpha(i, j) \neq \mathbf{0}$ for all edges $(i, j) \in \mathcal{E}$
2. $\alpha(i, j) = \mathbf{0}$ for all non-edges $(i, j) \notin \mathcal{E}$

where $i, j \in \mathcal{V}$ are distinct nodes. According to the above definition, graphs have the following properties:

1. Attributes $\alpha(i, i)$ of nodes $i \in \mathcal{V}$ may take any value from $\mathcal{A}$.
2. Graphs are complete, because non-edges are edges with zero attribute $\mathbf{0}$.
3. The definition comprises directed as well as undirected graphs.

Matrix representations. It is convenient to identify graphs with sets of matrices. A graph X is completely specified by a matrix representation $\boldsymbol{X} = (\boldsymbol{x}_{ij})$, where the elements $\boldsymbol{x}_{ij} \in \mathcal{A}$ represent the node and edges attributes for all $i, j \in \mathcal{V}$.

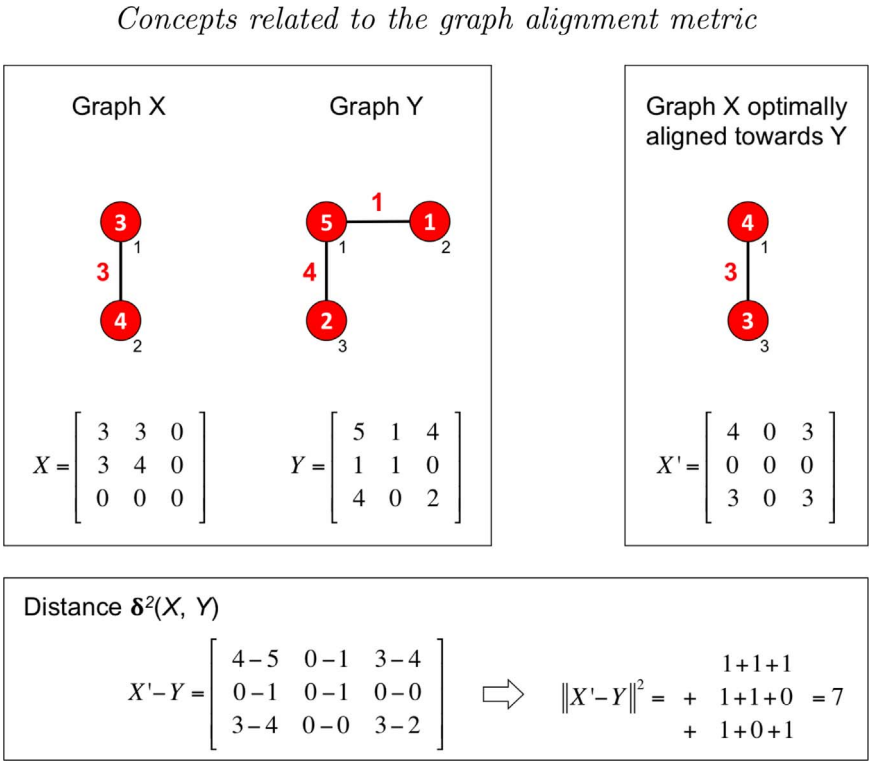

Fig. 1. The upper left box shows two attributed graphs X and Y, where attributes are weights. White numbers inside the nodes are the node attributes and red numbers attached to the edges are the edge attributes. Small black numbers next to the nodes are their unique identifiers. These identifiers correspond to the order of nodes according to the matrix representation given below for each graph. In this example, we assume that all graphs are of bounded order 3. Then all matrix representations have dimension 3×3. The matrix representation of the two-node graph X has a padding zero column and row. The box at the upper right shows the identifiers of the nodes of graph X when optimally aligned towards graph Y and its matrix representation $\boldsymbol{X}'$. The graph alignment metric is defined by $\delta(X, Y) = \left\|\boldsymbol{X}' - \boldsymbol{Y}\right\| = \sqrt{7}$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)