Iterative Optimum-Path Forest: A Graph-Based Data Clustering Framework

David Aparco-Cardenas
Institute of Computing
University of Campinas
Email: daparcoc@gmail.com

Alexandre X. Falcão
Institute of Computing
University of Campinas
Email: afalcao@ic.unicamp.br

Pedro J. de Rezende
Institute of Computing
University of Campinas
Email: rezende@ic.unicamp.br

***Abstract*—Data clustering is widely recognized as a fundamental technique of paramount importance in pattern recognition and data mining. It is extensively used in many fields of the sciences, business and engineering, covering a broad spectrum of applications. Despite the large number of clustering methods, only a few of them take advantage of optimum connectivity among samples for more effective clustering. In this work, we aim to fill this gap by introducing a novel graph-based data clustering framework, called Iterative Optimum-Path Forest (IOPF), that exploits optimum connectivity for the design of improved clustering methods. The IOPF framework consists of four fundamental components: (*i*) sampling of a seed set $\mathcal{S}$, (*ii*) partition of the graph induced from the dataset samples by an Optimum-Path Forest (OPF) rooted at $\mathcal{S}$, (*iii*) recomputation of $\mathcal{S}$ based on the previous graph partition, and, after multiple iterations of the last two steps, (*iv*) selection of the forest with the lowest total cost across all iterations. IOPF can be regarded as a generalization of the Iterative Spanning Forest (ISF) framework for superpixel segmentation from the image domain to the feature space. Herein, we present four IOPF-based clustering solutions to illustrate distinct choices of its constituent components. These are thereafter employed to address three different problems, namely, unsupervised object segmentation, road network analysis and clustering of synthetic two-dimensional datasets, in order to assess their effectiveness under various graph topologies, and to ascertain their efficacy and robustness when compared to competitive baselines.**

I. Introduction

A vast amount of data is generated by a wide range of sources in the current digital age. This data needs to be processed, analyzed, and transformed into valuable insights to support decision-making tasks. However, intensive computing resources and sophisticated techniques are required to efficiently and effectively extract the requisite information. Conventionally, these techniques are categorized into supervised, unsupervised and semi-supervised based on their reliance on labeled data. Supervised and semi-supervised methods depend, in different degrees, on labeled datasets, whose construction may become a time-consuming and tedious task in most situations. Unsupervised approaches can considerably alleviate this issue by not taking labels into consideration. Among the available unsupervised techniques, data clustering has become a crucial and widely used technique to discover hidden patterns and relationships in the data. It can considerably reduce the dependency on labeled data by assuming that samples in a cluster share the same label.

Clustering is a fundamental process that seeks to identify the intrinsic grouping in a set of unlabeled data based on some similarity measure. The goal of clustering is to partition a set of unlabeled objects into subsets (clusters) so that those within the same subset are more closely related (similar) to each other than to those falling in different subsets. Accordingly, designing practical clustering algorithms aiming to maximize both intra-subset similarity and inter-subset dissimilarity according to a similarity criterion remains a relevant research challenge. Clustering has a variety of applications in a broad range of domains, including plant and animal ecology, sequence analysis, human genetic clustering, medical imaging, market research, social network analysis, image segmentation, evolutionary algorithms, crime analysis, petroleum geology, physical geography, and so forth [1].

The most widely used clustering technique is *k-means*, a partitional clustering algorithm that stands out for its simplicity of implementation and intuitiveness. It is a numerical, non-deterministic, and iterative method that approximates each cluster's center by representing the objects as data points in the Euclidean space and measuring the dissimilarity between a pair of points by their Euclidean distance [2]. Despite being extensively used, *k-means* presents some shortcomings, such as that it can only identify spherical-shaped and symmetrical clusters [3]. Several extensions have been proposed to overcome these limitations [4], [5], which, however, address only a subset of these issues.

A. Objectives

In this context, the present work aims to explore graph-based clustering solutions for different applications through the proposal of a novel graph-based iterative clustering framework consisting of a sequence of *Optimum-Path Forest* (OPF) [6] executions. The objectives of this work are as follows: (*a*) formally present the *Iterative Optimum-Path Forest* (IOPF) framework, which is capable – through the different selection of its components – of creating a variety of clustering solutions that preserve connectivity among the samples of a dataset; (*b*) study and analyze the applications of IOPF-based solutions under different graph topologies while showcasing its flexibility, extensibility, and applicability to a wide range of problems; (*c*) analyze the effect of using IOPF with dynamic arc-weight estimation – an approach

that has proven its effectiveness for superpixel and object segmentation [7], [8], and whose application to the feature space is still unaddressed.

B. Contributions

Our main contribution is the proposal of a graph-based clustering framework, which, through a sequence of OPF executions, each followed by a seed recomputation stage, aims to partition a dataset while preserving connectivity within each cluster. A previous work [9] presented an algorithm with a similar formulation. However, it restricted its choice of components to the general *Image Foresting Transform* (IFT) [10] algorithm with the f_{sum} connectivity function. Thus, the inclusion of IFT with dynamic arc-weight estimation along with the $f_{\max}$ connectivity function as part of the set of framework components becomes also part of our contribution. Furthermore, we explore the effectiveness of IOPF-based solutions in various applications, which allows us to show the framework's flexibility under different graph configurations.

Our contributions also include the analysis of IOPF-based solutions under the following graph settings: (*a*) the adjacency relation and arc-weights are established from the problem definition; (*b*) only the adjacency relation comes from the problem definition; and (*c*) neither the adjacency relation nor the arc-weights are determined from the problem definition. Moreover, we propose strategies to build suitable graph topologies for the graph setting given in (*c*).

The interested reader is also referred to our paper [11] presented at the CIARP 2021 Conference (https://ciarp25.org/) and to the book chapter [12] published by Elsevier in 2022.

II. Related Work

Most of the more popular graph-based clustering algorithms do not exploit optimum connectivity between samples and seeds for cluster delineation. In this context, several OPF-based clustering algorithms have been introduced to bridge this gap, which can be broadly categorized into density-based and centroid-based algorithms.

Among the density-based techniques, Rocha *et al.* [13] introduced a first clustering method based on optimum connectivity – the maxima of a probability density function (pdf) compete among themselves to conquer the remaining samples of the dataset, and each maximum (dome of the pdf) defines a cluster as an optimum-path tree rooted on it. The pdf is estimated from a k-Nearest Neighbor (kNN) graph, and the choice of k is attained by finding the solution that minimizes a normalized graph-cut measure. Costa *et al.* [14] propose nature-inspired optimization techniques to speed up the selection of k for pdf estimation with application to intrusion detection in computer networks. Cappabianco *et al.* [15] extended the OPF-based clustering approach for large datasets by subsampling training samples, generating candidate solutions and selecting the most plausible one. The authors of the aforementioned work demonstrated the advantages of the method for MR-brain tissue segmentation. Montero and Falcão [16] propose a two-level divide-and-conquer clustering approach based on density-based OPF clustering. Such technique is well suited to handle large datasets. Chen *et al.* [17] presented an improved OPF-based clustering algorithm for segmentation of remote sensing images based on the principle that cluster centers display high local densities, whereas samples surrounding centers usually exhibit relatively low local densities. Afonso *et al.* [18] introduced a multi-layered OPF-based clustering algorithm inspired by hierarchical clustering. This algorithm, called *Deep Optimum-Path Forest*, builds a model comprised of a fixed number of stacked layers, such that the last layer contains the desired number of clusters. Recently, this algorithm was used in [19] to design visual dictionaries for the automatic identification of Parkinson's disease.

On the other hand, among the centroid-based techniques, Soor *et al.* [9] proposed *Iterated Watersheds* (IW), a graph-based clustering algorithm based on iterative applications of watershed transforms in a feature space from sets of enhanced cluster prototypes (seeds). This algorithm is a modified version of *k-means* with connectivity constraints, which, in turn, can be regarded as a particular configuration of the IOPF framework proposed herein.

III. Iterative Optimum-Path Forest

An IOPF-based method can essentially be summarized into four steps: (*i*) sampling of an initial seed set $\mathcal{S}$, (*ii*) graph partition by OPF from $\mathcal{S}$ into a graph derived from the dataset, (*iii*) recomputation of $\mathcal{S}$ based on the previous graph partition and, after multiple executions of steps (*ii*) and (*iii*), (*iv*) selection of the forest with the lowest total path-cost across all iterations.

Let $\mathcal{Z}$ be a dataset such that for every sample $s \in \mathcal{Z}$, there is a feature vector $v(s) \in \mathbb{R}^n$. For a given adjacency relation $\mathcal{A} \subseteq \mathcal{Z} \times \mathcal{Z}$, the pair $G = (\mathcal{Z}, \mathcal{A})$ defines a graph. The adjacency relation $\mathcal{A}$ can be defined in different ways, based on the specification of the problem at hand. In some cases, the adjacency relation of the graph is given beforehand, whereas in other situations it must be built from scratch. For instance, if $\mathcal{Z}$ is the set of pixels $s = (x_s, y_s)$ in the bi-dimensional domain of an image, $\mathcal{A}$ may be defined as $\mathcal{A}_r = \{(s,t) \in \mathcal{Z} \times \mathcal{Z} \mid 1 \leq \|(x_t, y_t) - (x_s, y_s)\| \leq r\}$. In this regard, the most notable adjacency relations on this domain are $\mathcal{A}_1$ and $\mathcal{A}_{\sqrt{2}}$, referred to as 4- and 8-neighborhood, respectively. As r increases, the local image feature space is explored with less spatial constraint. On the other hand, for arbitrary datasets, we may define $\mathcal{A}$ as follows:

1) $\mathcal{A} = \{(s,t) \in \mathcal{Z} \times \mathcal{Z} \mid s,t \in \mathcal{Z} \text{ and } s \neq t\}$, so that G represents a complete graph; or
2) $\mathcal{A} = \{(s,t) \in \mathcal{Z} \times \mathcal{Z} \mid v(t)$ is a k-nearest neighbor of $v(s)\}$, for a fixed k.

Nonetheless, in 2, it is essential to make sure that all nodes in $\mathcal{Z}$ are reachable from any seed in the seed set $\mathcal{S}$. Therefore, two conditions should be met: (*a*) if $(s,t) \in \mathcal{A}$, then $(t,s) \in \mathcal{A}$, and (*b*) G must be connected.

A *simple graph* with terminus t is a sequence of samples $\pi_t = \langle s_1, s_2, \ldots, s_n = t\rangle$,$(s_i, s_{i+1}) \in \mathcal{A}$, for $i \in \{1, 2, \ldots, n-1\}$, whereas $\pi_t = \langle t\rangle$ is called a *trivial path*. We

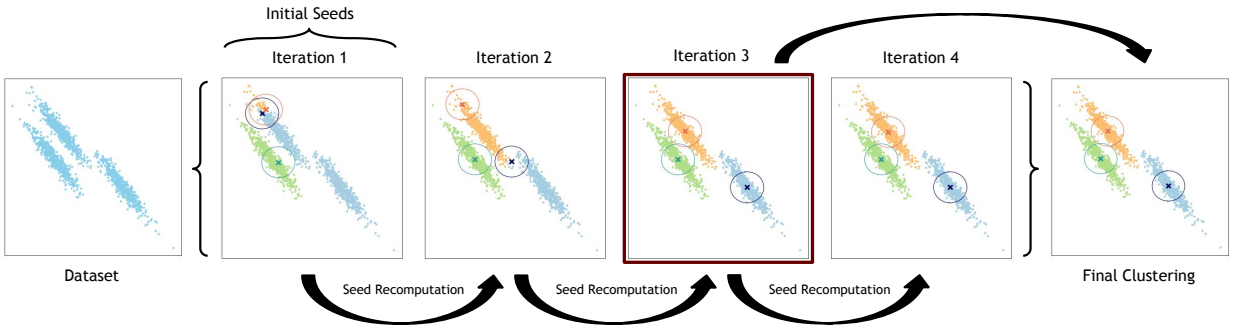


Fig. 1. IOPF pipeline. Initial seeds are selected randomly and recomputed at the end of each OPF execution. In this example, seed convergence is attained at the fourth iteration. Lastly, the partition that minimizes $\sum_{\forall s \in \mathcal{Z}} C(s)$ across all iterations is returned as the final clustering.

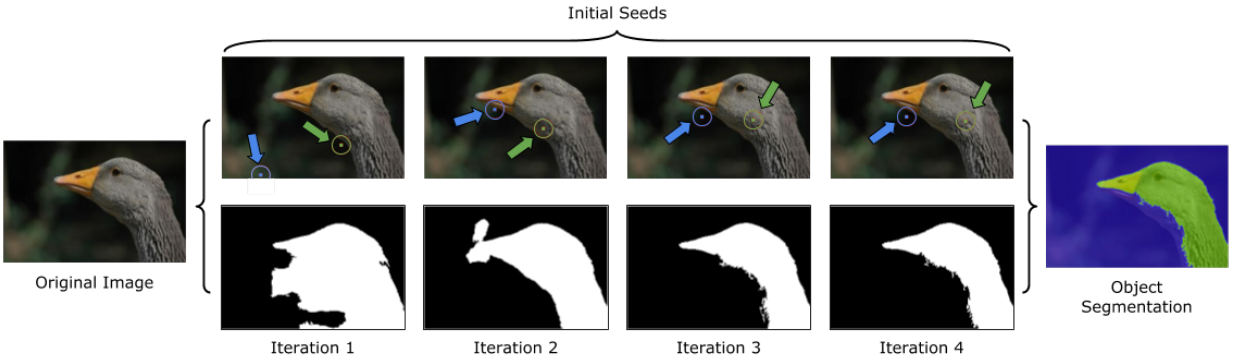


Fig. 2. Object delineation using the IDT algorithm with random initial seed set.

consider two types of connectivity functions, $f_{\max}$ and f_{sum}, with the same rule f_*, $* \in \{\max, \text{sum}\}$, for trivial paths:

$$\begin{aligned} f_*(\langle t \rangle)) &= \begin{cases} 0 & \text{if } t \in \mathcal{S} \subset \mathcal{Z} \\ +\infty & \text{otherwise} \end{cases} \\ f_{\max}(\pi_s \cdot \langle s,t \rangle) &= \max\{f_{\max}(\pi_s), w(s,t)\} \quad (1) \\ f_{\text{sum}}(\pi_s \cdot \langle s,t \rangle) &= f_{\text{sum}}(\pi_s) + w(s,t), \quad (2) \end{aligned}$$

where $w(s,t)$ is an arc-weight of $\langle s,t \rangle$ and $\pi_s \cdot \langle s,t \rangle$ denotes the concatenation of π_s and $\langle s,t \rangle$, with the two instances of s merged into one. The OPF algorithm minimizes a path-cost map $C(t) = \min_{\forall \pi_t \in \Pi_t}\{f_*(\pi_t)\}$, where Π_t is the set of all possible paths rooted at $\mathcal{S}$ with terminus t, while it outputs an *optimum-path forest* P – *i.e.*, an acyclic map that assigns to each $t \in \mathcal{Z}$ either its predecessor $P(t) \in \mathcal{Z}$ in the optimum path π_t^* rooted at $\mathcal{S}$ or a distinct marker *nil* if t is a *root* of the map (*i.e.*, $t \in \mathcal{S}$). Thus, each seed $t \in \mathcal{S}$ defines an optimum-path tree $\mathcal{T}_t$ (*i.e.*, cluster) in P, and may also propagate its corresponding label $L(t) \in \{1, 2, \ldots, k\}$ to its most strongly connected samples in $\mathcal{T}_t$.

An IOPF-based solution aims to estimate the graph partition that minimizes the total path-cost given by the sum of path costs between samples and their most strongly connected seeds in G. The minimization of this objective function is addressed following an iterative approach consisting in, given a fixed number of clusters k, partitioning the graph G into k optimum-path trees by multiple OPF executions from enhanced sets of seeds. Each OPF execution will output a triplet (L, C, P) consisting of a label map L, a cost map C, and a predecessor map P, leading to the computation of the total path-cost given by $\sum_{\forall s \in \mathcal{Z}} C(s)$. The set of enhanced seeds is recomputed at the end of each iteration, selecting the samples closest to each optimum-path tree's mean feature vector. The iterative procedure is repeated until either seed set convergence is achieved or a fixed maximum number of iterations is executed. Figure 1 depicts the pipeline of the IOPF framework, where initial seeds are randomly selected, and seed convergence is achieved at the fourth iteration. In this example, the third iteration minimizes $\sum_{\forall s \in \mathcal{Z}} C(s)$, and is, then, returned as the final clustering.

A. Iterative Dynamic Trees

Since IOPF is a generalization of the *Iterative Spanning Forest* (ISF) framework [20] from the image domain to the feature space, its application to image segmentation is straightforward. We call the methods for object delineation *Iterative Dynamic Trees* (IDT) [11]. A two-dimensional image is a pair $(\mathcal{D}_\mathcal{I}, \mathbf{I})$, such that $\mathbf{I}(p)$ assigns local image features (*e.g.*, color space components) for each pixel $p \in \mathcal{D}_\mathcal{I} \subset \mathbb{Z}^2$. An image can be rendered as a graph $(\mathcal{N}, \mathcal{A})$ under various configurations, depending on how the nodes $\mathcal{N} \subseteq \mathcal{D}_\mathcal{I}$ and the adjacency relation $\mathcal{A} \subset \mathcal{N} \times \mathcal{N}$ are defined. We define pixels

TABLE I
AMI, ARI, BOUNDARY RECALL AND CLUSTER ACCURACY (MEAN +/- STD. DEVIATION) FOR WEIZMANN 1-OBJECT AND 2-OBJECT DATASETS FOR IDT VARIANTS, DISF, IW-MAX AND IW-SUM.

	Method	AMI	ARI	BR	CA
1-Object	IDT_1	**0.564673 ± 0.283**	**0.613058 ± 0.317**	**0.657833 ± 0.241**	**0.908387 ± 0.091**
	IDT_2	0.344623 ± 0.270	0.363208 ± 0.323	0.433819 ± 0.276	0.841895 ± 0.114
	IDT_3	0.366932 ± 0.307	0.372370 ± 0.363	0.458131 ± 0.285	0.860064 ± 0.107
	DISF	0.304520 ± 0.282	0.282088 ± 0.347	0.398606 ± 0.296	0.836631 ± 0.112
	IW-max	0.397320 ± 0.278	0.419055 ± 0.318	0.473212 ± 0.276	0.856288 ± 0.112
	IW-sum	0.352781 ± 0.257	0.373990 ± 0.300	0.330048 ± 0.243	0.847699 ± 0.108
2-Object	IDT_1	**0.589247 ± 0.278**	0.600024 ± 0.345	**0.748527 ± 0.194**	**0.953605 ± 0.054**
	IDT_2	0.587252 ± 0.278	**0.614408 ± 0.333**	0.730065 ± 0.207	0.946522 ± 0.064
	IDT_3	0.386087 ± 0.279	0.334149 ± 0.328	0.518125 ± 0.263	0.902305 ± 0.100
	DISF	0.420036 ± 0.295	0.376453 ± 0.352	0.582483 ± 0.263	0.919615 ± 0.078
	IW-max	0.435559 ± 0.330	0.544933 ± 0.311	0.615948 ± 0.231	0.921671 ± 0.086
	IW-sum	0.395757 ± 0.242	0.347743 ± 0.299	0.496769 ± 0.224	0.895421 ± 0.097

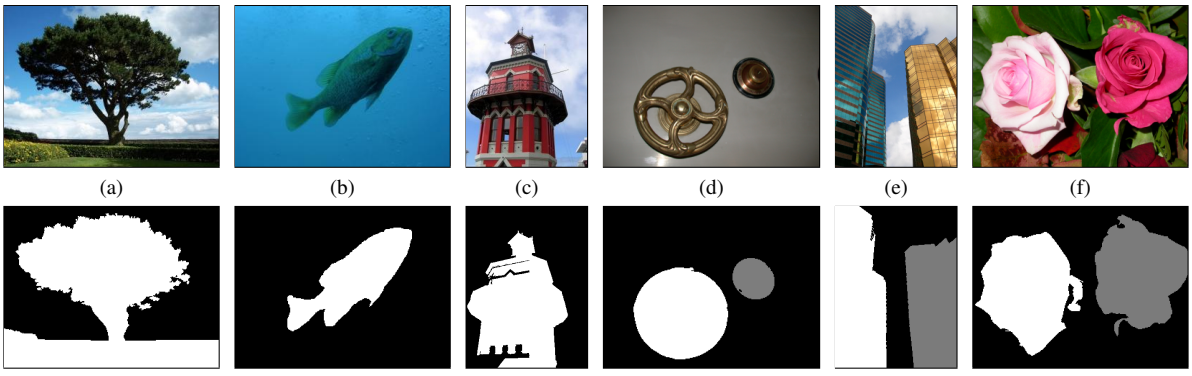


Fig. 3. Segmentation results for Weizmann 1-Object and 2-Object datasets. (a)–(f) Original images, (g)–(l) IDT_1.

as nodes ($\mathcal{N} = \mathcal{D}_\mathcal{I}$), such that $\mathbf{I}(p)$ represents the CIELab color components of pixel p, and the 8-neighborhood relation defines the arcs.

All the framework components presented in the previous sections remain valid for this application. Nevertheless, given the nature of the problem, some other strategies can be introduced as components for the framework. Following this line of thought, we present a new seed recomputation strategy in the image domain. During iteration j, new seeds are selected as the nodes closest to the mean feature vector for each optimum-path tree $\mathcal{T}_{ij}, i \in \{1, 2, \ldots, k\}$. Nonetheless, in the image domain, we may also select the new seeds as the nodes closest to the mean pixel of each optimum-path tree. The mean pixel is defined as the arithmetic mean of pixel coordinates of the elements of clusters $\mathcal{C}_{ij}, i \in \{1, 2, \ldots, k\}$. Thus, each seed $r_{i,j+1}$, for $i = 1, 2, \ldots, n$, for iteration $j+1$ is computed as:

$$r_{i,j+1} = \underset{p \in \mathcal{C}_i}{\operatorname{argmin}} \{ \| p - \frac{1}{|\mathcal{C}_i|} \times \sum_{\forall q \in \mathcal{C}_i} q \| \}. \quad (3)$$

Figure 2 illustrates the application of the IDT algorithm to a natural image from a random initial seed set.

IV. EXPERIMENTAL RESULTS

In this section, we present three applications to illustrate the robustness and flexibility of the IOPF framework by designing suitable and effective IOPF-based methods for each problem's context.

A. Object Delineation by Iterative Dynamic Trees

To demonstrate the advantages of step (*iv*) and random seed sampling in step (*i*), we compare three versions of IDT. IDT_1 is the proposed version, as described in the previous section. IDT_2 is IDT_1 without step (*iv*), selecting the optimum-path forest of the last iteration, as commonly adopted by ISF-based methods, such as DISF [8]. IDT_3 is IDT_1 using grid sampling in step (*i*), as used by DISF and most superpixel segmentation methods.

To demonstrate the improvement of IDT for object delineation, we compare it against DISF and IW [9] with two path-cost functions: IW-max computes the cost of a path as the maximum arc weight along it, for fixed arc weights $\|\mathbf{I}(q) - \mathbf{I}(p)\|$, whereas IW-sum computes the cost of a path as the sum of its arc weights. DISF begins from a set of 150 seeds selected by grid sampling for all images, and reduces the seed set size in every iteration until it reaches the number

of desired objects. IW has already been demonstrated to be superior to spectral clustering, isoperimetric partitioning and k-means for the task of object delineation [9].

For evaluation of object segmentation, we use the Weizmann 1-Object and 2-Object datasets [21], containing 100 images each, along with ground-truth segmentations. Images in these datasets (available at http://www.wisdom.weizmann.ac.il/~vision/Seg_Evaluation_DB/) depict one or two objects in the foreground. Table I shows the effectiveness of object segmentation for all methods according to four different metrics (AMI, ARI, BR, and CA). IDT_1 stands out as the best approach, obtaining a better border delineation and accuracy than its counterparts. Moreover, it can also be stated that random sampling suffices for step (*i*), and step (*iv*) included by the proposed approach into the ISF framework improves object segmentation.

B. Analysis on Road Networks

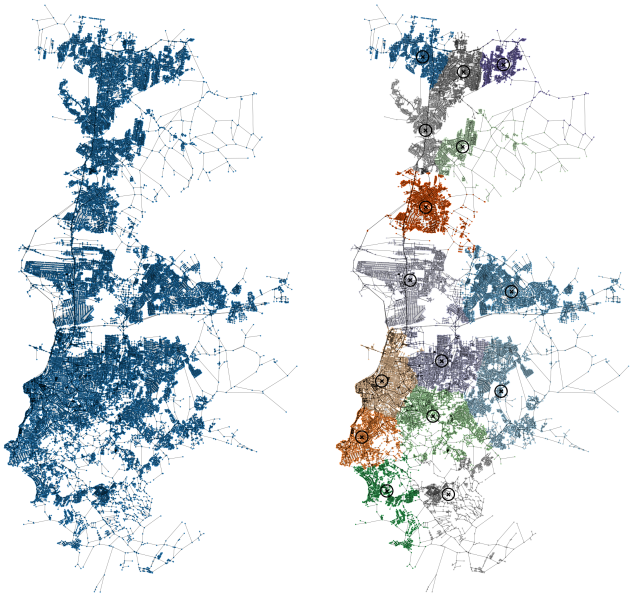
The adjacency relation in road networks is defined beforehand by the road network map, where edges, represented by roads, connect a pair of reference points.

The problem this experiment addresses is described as follows: given a road network instance, identifying appropriate points for placing emergency stations, such that the emergency station reaches the point of an incident in the minimum time possible. The devised solution must comply with the following constraints: (*a*) The emergency station must be reachable from the point of an incident in a short time interval (*i.e.*, the distance between these two points must be minimized), and (*b*) the number of emergency stations layed out across the map must be as low as possible to reduce the establishment costs. A road network will induce a weighted graph $G = (\mathcal{Z}, \mathcal{A})$, where the nodes are defined by a set of reference points spread across the road map. Such points, uniquely identified by a pair of coordinates $x = (x_1, x_2)$, constitute the dataset $\mathcal{Z}$. An emergency station is established at one of such points. The adjacency relation $\mathcal{A}$ that defines the arcs of the graph is given by a set of pairs $(x, y) \in \mathcal{Z} \times \mathcal{Z}$, such that x and y are connected by a road. The arcs are weighted by their corresponding road lengths, which are provided in advance for the experiment.

Based on the above definition, this problem may be formulated as that of discovering a set of k emergency points $c_i \in \mathcal{Z}, i \in \{1, 2, \ldots, k\}$, such that the sum of path-costs between each reference point $s \in \mathcal{Z}$ and its closest emergency station – *i.e.*, $\sum_{s \in \mathcal{Z}} f(\pi_s)$ – is minimized across all reference points, for a given connectivity function f. As it happens, the application of the IOPF framework to this problem is straightforward. In this context, the problem described above can be divided into two subproblems: (*a*) discovering the set of k emergency stations through the IOPF framework, and (*b*) computing the sum of path-costs between each of those emergency stations and its closest points. After repeating the experiment with a sequence of increasing values, the ideal number of stations is determined so that the reduction of $\sum_{s \in \mathcal{Z}} f(\pi_s)$ does not compensate the placing cost for establishing an additional station.

The road networks for this experiment were obtained from [22] (available at https://figshare.com/articles/dataset/Urban_Road_Network_Data/2061897). In this experiment, our objective is to determine the IOPF configuration that suits the problem described above. Hence, we compare four versions of IOPF, namely, IW-sum and IW-max using f_{sum} and $f_{\max}$ with fixed arc weights $\|x - y\|$, and IOPF-dynsum and IOPF-dynmax using f_{sum} and $f_{\max}$ with dynamic arc-weight estimation. In [9], a similar experiment was conducted where IW exhibited better performance than k-means and greedy k-center. This experiment uses the road networks corresponding to the Brazilian cities of São Paulo, Rio de Janeiro, Belo Horizonte, Recife, Porto Alegre, and Salvador.

Figure 4 (a) shows the graph induced by the road networks of the city of Recife, where blue dots represent the nodes or reference points, while black lines linking pairs of reference points represent the edges or roads. Figure 4 (b) shows the clustering result of IW-sum with 15 centers for each city's network, where each cluster is colored with a different color and the centers (*i.e.*, emergency points) are marked with an encircled point.



(a) (b)

Fig. 4. Road map network of the city of Recife, Brazil in (a) and the clustering result of IW-sum for 15 emergency stations in (b).

The experiments were carried out using the same sets of initial seeds for the four clustering solutions. For the road network of every one of the cities, each method was executed thirty times for a varying number of centers. Next, the sum of path-costs across all nodes $\sum_{s \in \mathcal{Z}} f(\pi_s)$ was averaged across all executions to assess their effectiveness. From the results[1], it

[1]The reader will find the tables and figures corresponding to the complete set of results in text of the master's thesis [23].

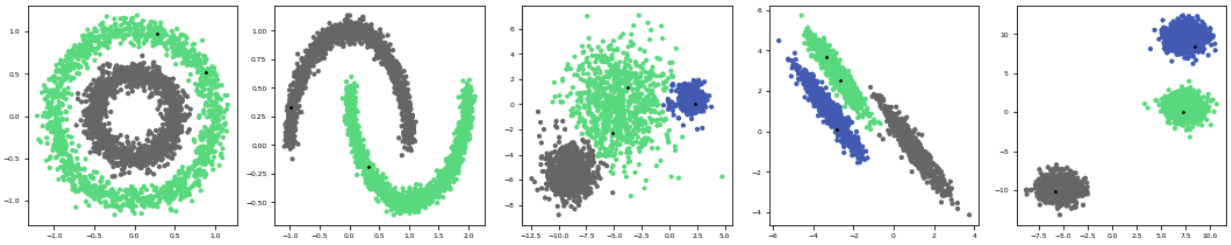


Fig. 5. Clustering results on the synthetic datasets using IW-max for the complete-graph topology.

can be concluded that IW-sum is the most suitable IOPF-based method in most cases, since it achieves a lower value than its counterparts and is worse than IOPF-dynsum and IOPF-dynmax in only a few cases where the difference in values is not significant.

C. Experiments on Synthetic Datasets

In order to verify the performance and robustness of the IOPF framework in a wider variety of datasets, we evaluate several IOPF-based methods on synthetic datasets that exhibit a broad spectrum of shapes and distributions. In such datasets, we do not possess enough information regarding the underlying relationship among the samples to establish a suitable graph topology, as opposed to what we considered in Sections IV-A and IV-B.

Let $\mathcal{Z}$ be a dataset such that each sample $s \in \mathcal{Z}$ is represented in the feature space by a feature vector $v(s) \in \mathbb{R}^n$. The adjacency relation $\mathcal{A} \subseteq \mathcal{Z} \times \mathcal{Z}$ may be defined in such a way that the induced graph $G = (\mathcal{Z}, \mathcal{A})$ can be established either as a complete or a k-nearest neighbor graph.

We conducted experiments with four configurations of IOPF, namely, IW-sum, IW-max, IOPF-dynsum and IOPF-dynmax, to determine their effectiveness under the complete graph topology. We compared the four methods with k-means to assess their performance against the most widely used clustering algorithm. In this setting, IW-max is the IOPF-based method that best separates the groups for all synthetic datasets among all those tested. Figure 5 shows the results of this experiment for the IW-max method. Alternatively, we may also define G as a k-nearest neighbor graph, where each sample $s \in \mathcal{Z}$ is linked through an edge to its k closest neighbors for a fixed k. From the results[2], we conclude that IW-sum is now able to successfully separate all the groups, achieving the same performance as that of IW-max. Therefore, imposing restrictions on the graph topology leads to improvements in the clustering capabilities of IW-sum.

To ascertain the framework's effectiveness and robustness against other state-of-the-art clustering algorithms, we compared IW-max using a complete graph topology, and a seed selection algorithm consisting in a sequence of OPF executions, described in more detail in the master's thesis text [23], against five popular clustering algorithms: (*a*) mean shift, (*b*) spectral clustering, (*c*) DBSCAN, (*d*) Gaussian mixture, and (*e*) agglomerative clustering. We are able to conclude[2] that, in contrast to its counterparts, IW-max successfully separates the groups for all synthetic datasets. To further assess the framework's performance, we employed seven additional datasets obtained from [24]. Each dataset presents its own challenges due to their inherent structure and distribution. Once again, the IW-max method is the only method to correctly identify all the groups for all datasets.

[2]Again, the reader is referred to text of the master's thesis [23] to see all tables and figures corresponding to the complete set of results.

V. Conclusion

We introduced a flexible and robust graph-based clustering framework, called Iterative Optimum-Path Forest (IOPF), that employs consecutive executions of the OPF algorithm from re-estimated seed sets to partition an input dataset, while allowing the design of connectivity-based clustering methods by suitable choices of its components. Moreover, we introduced an algorithm to select initial seeds for data clustering, improving the results presented in [9]. In this context, we described four IOPF-based clustering methods, IW-sum, IW-max, IOPF-dynsum, and IOPF-dynmax. We evaluated them for object delineation, identification of emergency stations in road networks, and group identification in synthetic two-dimensional datasets with different shapes and sizes. We observed that IW-sum improves effectiveness when the graph topology is constrained to the k-nearest neighbors. On the other hand, IOPF-dynmax, previously called IDT [11], appears as the best approach for object delineation, while IW-sum is the best suited for identification of emergency stations in road networks, and IW-max is the winner for clustering of two-dimensional datasets.

We intend to investigate new techniques for seed recomputation in order to further improve the effectiveness of the IOPF-based methods, to include local density information in the identification of initial seeds, and to explore new applications for the IOPF-based methods.

Acknowledgment

This research was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Brazilian National Council for Scientific and Technological Development (CNPq), #313329/2020-6, #309627/2017-6, #303808/2018-7, #140930/2021-3, São Paulo Research Foundation (FAPESP), #2020/09691-0, #2018/26434-0, #2014/12236-1.

REFERENCES

[1] K. S. Dar, I. Javed, W. Amjad, S. Aslam, and A. Shamim, "Survey of clustering applications," *Journal of Network Communications and Emerging Technologies (JNCET)*, vol. 4, no. 3, 2015.

[2] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[3] J. P. Ortega, M. Del, R. B. Rojas, and M. J. Somodevilla, "Research issues on k-means algorithm: An experimental trial using matlab," in *CEUR workshop proceedings: semantic web and new technologies*, 2009, pp. 83–96.

[4] D. Pelleg, A. W. Moore *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters." in *Icml*, vol. 1, 2000, pp. 727–734.

[5] R. M. Esteves, T. Hacker, and C. Rong, "Competitive k-means, a new accurate and distributed k-means algorithm for large datasets," in *2013 IEEE 5th International Conference on Cloud Computing Technology and Science*, vol. 1, 2013, pp. 17–24.

[6] J. P. Papa, A. X. Falcão, and C. T. N. Suzuki, "Supervised pattern classification based on optimum-path forest," *International Journal of Imaging Systems and Technology*, vol. 19, no. 2, pp. 120–131, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.20188

[7] J. Bragantini, S. B. Martins, C. Castelo-Fernandez, and A. X. Falcão, "Graph-based image segmentation using dynamic trees," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2018, pp. 470–478.

[8] F. C. Belém, S. J. F. Guimarães, and A. X. Falcão, "Superpixel segmentation using dynamic and iterative spanning forest," *IEEE Signal Processing Letters*, vol. 27, pp. 1440–1444, 2020.

[9] S. Soor, A. Challa, S. Danda, B. Daya Sagar, and L. Najman, "Iterated watersheds, a connected variation of k-means for clustering gis data," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–1, 2019.

[10] A. X. Falcão, J. Stolfi, and R. de Alencar Lotufo, "The image foresting transform: theory, algorithms, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 19–29, 2004.

[11] D. Aparco-Cardenas, P. J. de Rezende, and A. X. Falcão, "Object delineation by iterative dynamic trees," in *Iberoamerican Congress on Pattern Recognition*, 2021, pp. 1–10.

[12] D. Aparco-Cardenas, P. J. de Rezende, and A. X. Falcão, "Chapter 8 - An iterative optimum-path forest framework for clustering," in *Optimum-Path Forest*, A. X. Falcão and J. P. Papa, Eds. Academic Press, 2022, pp. 175–216. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128226889000165

[20] J. E. Vargas-Muñoz, A. S. Chowdhury, E. B. Alexandre, F. L. Galvão, P. A. Vechiatto Miranda, and A. X. Falcão, "An iterative spanning forest

[13] L. M. Rocha, F. A. Cappabianco, and A. X. Falcão, "Data clustering as an optimum-path forest problem with applications in image analysis," *International Journal of Imaging Systems and Technology*, vol. 19, no. 2, pp. 50–68, 2009.

[14] K. A. Costa, L. A. Pereira, R. Y. Nakamura, C. R. Pereira, J. P. Papa, and A. X. Falcão, "A nature-inspired approach to speed up optimum-path forest clustering and its application to intrusion detection in computer networks," *Information Sciences*, vol. 294, pp. 95–108, 2015.

[15] F. A. Cappabianco, A. X. Falcão, C. L. Yasuda, and J. K. Udupa, "Brain tissue mr-image segmentation via optimum-path forest clustering," *Computer Vision and Image Understanding*, vol. 116, no. 10, pp. 1047–1059, 2012.

[16] A. Echemendía Montero and A. X. Falcão, "A divide-and-conquer clustering approach based on optimum-path forest," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2018, pp. 416–423.

[17] S. Chen, T. Sun, F. Yang, H. Sun, and Y. Guan, "An improved optimum-path forest clustering algorithm for remote sensing image segmentation," *Computers & Geosciences*, vol. 112, pp. 38–46, 2018.

[18] L. Afonso, A. Vidal, M. Kuroda, A. X. Falcão, and J. P. Papa, "Learning to classify seismic images with deep optimum-path forest," in *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2016, pp. 401–407.

[19] L. C. Afonso, C. R. Pereira, S. A. Weber, C. Hook, A. X. Falcão, and J. P. Papa, "Hierarchical learning using deep optimum-path forest," *Journal of Visual Communication and Image Representation*, vol. 71, p. 102823, 2020.
framework for superpixel segmentation," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3477–3489, 2019.

[21] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2007.

[22] A. Karduni, A. Kermanshah, and S. Derrible, "A protocol to convert spatial polyline data to network formats and applications to world urban road networks," *Scientific data*, vol. 3, no. 1, pp. 1–7, 2016.

[23] D. Aparco-Cardenas, "Floresta de Caminhos Ótimos Iterativa: Um Arcabouço para Agrupamento de Dados Baseado em Grafos. MSc thesis, Universidade Estadual de Campinas," Master's thesis, Universidade Estadual de Campinas, 2021.

[24] P. Fränti and S. Sieranoja, "Clustering basic benchmark," http://cs.joensuu.fi/sipu/datasets/, accessed: 2020-09-30.