Pattern Recognition 46 (2013) 3371–3382

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Clustering and outlier detection using isoperimetric number of trees ☆

A. Daneshgar [a,*], R. Javadi [b], S.B. Shariat Razavi [a]

[a] Department of Mathematical Sciences, Sharif University of Technology, P.O. Box 11155–9415, Tehran, Iran
[b] Department of Mathematical Sciences, Isfahan University of Technology, P.O. Box 84156–83111, Isfahan, Iran

ARTICLE INFO

Article history:
Received 31 July 2012
Received in revised form
26 April 2013
Accepted 5 May 2013
Available online 23 May 2013

Keywords:
Isoperimetric constant
Cheeger constant
Normalized cut
Graph partitioning
Perceptual grouping
Data clustering
Outlier detection

ABSTRACT

We propose a graph-based data clustering algorithm which is based on exact clustering of a minimum spanning tree in terms of a minimum isoperimetry criteria. We show that our basic clustering algorithm runs in $O(n \log n)$ and with post-processing in almost $O(n \log n)$ (average case) and $O(n^2)$ (worst case) time where n is the size of the data-set. It is also shown that our generalized graph model, which also allows the use of potentials at vertices, can be used to extract an extra piece of information related to anomalous data patterns and outliers. In this regard, we propose an algorithm that extracts outliers in parallel to data clustering. We also provide a comparative performance analysis of our algorithms with other related ones and we show that they behave quite effectively on hard synthetic data-sets as well as real-world benchmarks.

1. Introduction

1.1. A concise survey of main results

Data clustering, as the unsupervised grouping of similar patterns into clusters, is a central problem in engineering disciplines and applied sciences which is also constantly under theoretical and practical development and verification. In this article we are concerned with graph based data clustering methods which are extensively studied and developed mainly because of their simple implementation and acceptable efficiency in a number of different fields such as signal and image processing, computer vision, computational biology, machine learning and networking to name a few.

The main contribution in this article can be described as a general graph-based data clustering algorithm which falls into the category of such algorithms that use a properly defined sparsest cut problem as the clustering criteria. In this regard, it is instructive to note some highlights of our approach before we delve into the details in subsequent sections (details of our approach as well as a survey of related contributions will appear in the second part of this introduction).

It has been already verified that graph-based clustering methods that operate in terms of non-normalized cuts are not suitable for general data clustering and behave poorly in comparison to the normalized versions (e.g. see [1]). Moreover, it is well known that there is a close relationship between the minimizers of the normalized cut problem, spectral clustering solutions, mixing rates of random walks, the minimizers of the K-means cost function, kernel PCA and low dimensional embedding, while the corresponding decision problems are known to be NP-complete in general (e.g. see [2–8] and references therein).

In this article, we will provide an efficient clustering algorithm which is based on a relaxation of the feasible space of solutions from the set of *partitions* to the larger set of *subpartitions* (i.e. mutually disjoint subsets of the domain). From one point of view, our algorithm can be considered as a generalization of Grady and Schwartz approach [9,10] based on isoperimetry problems while we extensively rely on the results of [3,4]. Also, we believe that this relaxation which is based on moving from the space of partitions to the space of subpartitions not only provides a chance of making the problem easier to solve but also is in coherence with the natural phenomena of having undesirable data or outliers. We will use this property to show that our algorithm can be enhanced to a more advanced procedure which is capable of presenting a hierarchy of data similarity profile which in turn can lead to the extraction of outliers.

In this regard, one may comment on some different aspects of this approach as follows.

Theoretical aspects: From a theoretical point of view, it is proved in [4] that the normalized cut criteria is not formally well defined in the sense that it does not admit a variational description through a real function relaxation of the problem (i.e. it does not

☆A preliminary version of this article has already been made available at arXiv:1203.4204.

* Corresponding author. Tel.: +982166165610; fax: +982166005117.
E-mail addresses: daneshgar@sharif.ir (A. Daneshgar), rjavadi@cc.iut.ac.ir (R. Javadi), basirshariat@alum.sharif.ir (S.B. Shariat Razavi).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.05.015

admit a Federer–Fleming type theorem). However, for $k\geq 2$, the well-defined version, known as the *k-isoperimetry problem* (defined in [4]), whose definition is in terms of normalized-flow minimization on k-subpartitions, actually admits such a relaxation. It should be noted that although there are some approaches to clustering which are based on the classical 2-isoperimetry (i.e. Cheeger constant) on weighted graphs (e.g. see [9]), but as it follows from the results of [4], in the classical model the difference between the cases of partitions and subpartitions only is observable when $k\geq 3$, and consequently, our approach is completely different in nature from iterative 2-partitioning or spectral approximation methods based on eigenmaps already existing in the literature.

Also, as a bit of a surprise (see Theorem 2), it turns out that a special version of the k-isoperimetry problem is *efficiently solvable* for weighted trees. This fact along with a well-known approach of finding an approximate graph partitioning through minimum spanning trees constitute the core of our algorithm.

Practical aspects: There are different practical aspects of the proposed algorithm that one may comment on. First, the proposed approximation algorithm run-time is almost linear in terms of the size of input-data which provides an opportunity to cluster large data sets. Also, it should be noted that our algorithm for k-clustering obtains an exact optimal clustering of a suitably chosen subtree in a global approach and does not apply an iterative two-partitioning or an approximation through eigenmaps. This in a way is one of the reasons supporting a better approximation of our algorithm compared to the other existing ones. In this regard, we also present a number of experimental results justifying a better performance of our algorithm in practice (see Section 3).

Second, we should note that approximation through the isoperimetry criteria provides an extra piece of information as a (possibly nonempty) subset of the domain (since the union of subpartitions may not be a covering). This piece of information makes it possible to obtain the almost minimal clustering as well as to extract deviated data and outliers, *at the same time* (see Section 4). In order to handle this extra information, we have generalized our graph model to the case of a *weighted graph with potential*. This generalization of the graph representation model is another original aspect of our contribution where we rely on results of [3,4] in this more general setting (see Theorem 2). We also provide comparative experimental results to analyse the efficiency of the proposed outlier detection method.

1.2. Background and related contributions

Unsupervised grouping of data based on a predefined similarity criteria is usually referred to as *data clustering* in general, where in some more specific applications one may encounter some other terms as *segmentation* in image processing or *grouping* in data mining. Based on its importance and applicability, there exists a very vast literature related to this subject (e.g. see [11–13] for some general background), however, in this article we are mainly concerned with clustering algorithms that rely on a representation of data as a simple weighted graph in which the edge-weights are tuned, using a predefined similarity measure (e.g. see Section 2, [14] and references therein).

Graph-based data clustering is usually reduced to the *graph partitioning* problem on the corresponding weighted graph which is also well-studied in the literature. To this end, it is instructive to note that from this point of view and if one considers a weighted graph as a geometric object, then the partitioning problem can be linked to a couple of very central and extensively studied problems in geometry as *isoperimetry problem*, *concentration of measure* and *estimation of diffusion rates* (e.g. see [4,5] and references therein).

A graph-based clustering or a graph partitioning problem is usually reduced to an optimization problem where the cost function is a measure of sparsity or density related to the corresponding classes of data. From this point of view, it is not a surprise to see a variety of such measures in the literature, however, from a more theoretical standpoint such similarity measures are well-studied and, at least, the most geometrically important classes of them are characterized (e.g. see [15] for a very general setting). In this context such measures usually appear as *norms* or their normalized versions that should be minimized or maximized to lead to the expected answer.

What is commonly refereed to as *spectral clustering* is the case in which the corresponding normalized norm is expressed as an L^2 (i.e. Euclidean) norm and admits a real-function relaxation whose minimum is actually an eigenvalue of the weight (or a related) matrix of the graph. This special case along with the important fact that, the spectral properties (i.e. eigenvalues and eigenfunctions) of a finite matrix can be effectively (at most in $O(n^3)$ time) computed, provides a very interesting setting for data clustering in which the corresponding optimization problem can be tackled with using the well-known tools of linear algebra and operator theory (e.g. see [16–24] and references therein for a general background in spectral methods).

Although, applying spectral methods are quite effective and vastly applied in data clustering, but still the time complexity of the known algorithms and also the approximation factor of this approach in not as good as one expects when one is dealing with large data-sets (e.g. see [25] for a recent algorithm and references therein). On the other way round, these facts lead one to consider the original normalized versions of the L^1 norm that reduces clustering to the sparsest (or similar minimal) cut problems or their real-function relaxations as the corresponding approximations. It is proved in [4] that the most natural such normalized norms *do not* admit real-function relaxations when they are minimized over partitions of their domain. Moreover, it is shown in the same reference that such normalized norms *do admit* real-function relaxations when they are minimized over subpartitions of their domain. In this new setting the minimum values, that correspond to the eigenvalues in the spectral L^2 setting, are usually referred to as *isoperimetric constants*.

Unfortunately, contrary to the case of L^2, decision problems corresponding to the isoperimetry problems are usually NP-hard (e.g. see [1,3,6,26]), which shows that computing the exact value of the isoperimetric constants is not an easy task. There has been a number of contributions in the literature whose main objectives can be described as proposing different methods to get around this hard problem and find an approximation for the corresponding isoperimetry problem as a criteria of clustering, and consequently, obtaining an approximate clustering of the given data.

In this regard, one may at least note two different approaches as follows. In the one hand, there have been contributions which have tried to reduce the problem to the more tractable case of trees by first finding a suitable subtree of the graph and then try to approximately cluster the tree itself (e.g. see [27–33]). The difference between such contributions usually falls into the way of choosing the subtree and the method of their clustering. On the other hand, one may also try to obtain a global clustering by a mimic of spectral methods through solving not an eigenfunction problem but a similar problem in L^1 (e.g. see [9,10]). These methods usually follow an iterative 2-partitioning since there was not much information about approximations for higher order eigenfunctions or similar solutions in L^1 until recently (e.g. see [6,9,25]).

Our main contribution in this article can be described as a culmination of above mentioned ideas that strongly rely on some recent studies of higher order solutions of isoperimetry problems (see [3,4]), in which we first search for a suitable spanning subtree and after that we obtain the *exact* solution of the corresponding optimization problem for our suitably chosen isoperimetric constant (see Section 2). Also, we will obtain a subset of data given as