Information Sciences 251 (2013) 96–113

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Clustering with a new distance measure based on a dual-rooted tree

Laurent Galluccio [a], Olivier Michel [b], Pierre Comon [b,*], Mark Kliger [c], Alfred O. Hero [d]

[a] Lab. Lagrange, Observatoire de la Cote d'Azur, BP.4229, 06304 Nice Cedex 4, France
[b] Gipsa-Lab UMR 5216, 961 Rue de la Houille Blanche, BP.46, 38402 Saint Martin d'Hères Cedex, France
[c] Omek Interactive Ltd., 2 Hahar Str, Har Tuv A, Bet Shemesh 99067, Israel
[d] Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109-2122, USA

ARTICLE INFO

Article history:
Received 28 August 2012
Received in revised form 15 February 2013
Accepted 30 May 2013
Available online 27 June 2013

Keywords:
Non-metric clustering
Minimal spanning tree
Prim's algorithm
Affinity measure
Co-association measure
Consensus clustering

ABSTRACT

This paper introduces a novel distance measure for clustering high dimensional data based on the hitting time of two Minimal Spanning Trees (MST) grown sequentially from a pair of points by Prim's algorithm. When the proposed measure is used in conjunction with spectral clustering, we obtain a powerful clustering algorithm that is able to separate neighboring non-convex shaped clusters and to account for local as well as global geometric features of the data set. Remarkably, the new distance measure is a true metric even if the Prim algorithm uses a non-metric dissimilarity measure to compute the edges of the MST. This metric property brings added flexibility to the proposed method. In particular, the method is applied to clustering non Euclidean quantities, such as probability distributions or spectra, using the Kullback–Leibler divergence as a base measure. We reduce computational complexity by applying consensus clustering to a small ensemble of dual rooted MSTs. We show that the resultant consensus spectral clustering with dual rooted MST is competitive with other clustering methods, both in terms of clustering performance and computational complexity. We illustrate the proposed clustering algorithm on public domain benchmark data for which the ground truth is known, on one hand, and on real-world astrophysical data on the other hand.

1. Introduction

The process of clustering partitions a set of data into non-overlapping subsets. The partitions are determined such that patterns belonging to the same cluster share more similarity with each other than with patterns belonging to different clusters [30]. Such problems have been investigated in many fields of research including: data mining [6], pattern recognition [46], image segmentation [50], computer vision [41] and bio-informatics [52]. There are a wide range of clustering methods available, e.g., hierarchical clustering, spectral clustering, graph partitioning algorithms and k-means [26,31,25,46]. In this paper we introduce a new clustering method that uses dual rooted trees combined with consensus methods. The approach is closely related to level-set methods [44] and entropy minimization [27]. However, dual rooted trees have advantageous mathematical properties and their performance is competitive with the state-of-the art.

Dissimilarity measures between data points play a crucial role in designing clustering algorithms. These measures determine how the clustering algorithm differentiates pairs of points within the same cluster (high similarity) from pairs of points

* Corresponding author. Tel.: +33 4 7682 6271; fax: +33 4 7657 4790.
E-mail addresses: laurent.galluccio@gmail.com (L. Galluccio), olivier.michel@gipsa-lab.inpg.fr (O. Michel), pierre.comon@grenoble-inp.fr (P. Comon), mark@medasense.com (M. Kliger), hero@umich.edu (A.O. Hero).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.05.040

in different clusters (low similarity). In many cases using Euclidean metric to measure dissimilarities between data points is insufficient. This has motivated spectral diffusion methods of clustering [42,37,34,51]. The original spectral method used a Gaussian kernel on a Euclidean metric to construct a more discriminating dissimilarity measure [37]. In [34], the Gaussian kernel was interpreted as a heat diffusion kernel which induces a random walk on the graph with nodes consisting of data points, yielding a measure of dissimilarity. In [39] a "commute time" dissimilarity measure is introduced, and is closely related to diffusion distance. In both of these approaches the diffusion and commute time dissimilarity measures are used for embedding data points into a new system of coordinates defined by the eigenvectors of the heat kernelized affinity matrix. The final clustering step is achieved by using k-means on the embedded data.

The dual rooted minimal spanning tree (MST) clustering approach proposed in this paper is different. Starting from a base dissimilarity measure between data points, it constructs MSTs rooted at different points in the dataset. It then defines the dissimilarity between pairs of points as the time it takes before collision of the two MSTs as they are grown from each root using Prim's algorithm [38]. This time is called the dual rooted tree hitting time, and it is a non-Euclidean dissimilarity measure that describes global as well as local geometrical properties of the data set, as explained in details in Section 2.2. In particular, this *hitting time* can be used as a measure of dissimilarity between the two roots and it is influenced by the distance between the roots in addition to the dissimilarity of their local neighborhoods. The matrix of pairwise dissimilarities can then be transformed into an affinity matrix by applying the standard heat kernel approach used in spectral clustering. This principal role of the local neighborhoods of each pair of points is one of the main differences between the dual rooted MST approach and the diffusion kernel and commute time methods.

The starting point for this paper is the simple algorithm described above, called the Symmetric Dual Rooted Prim Tree (SDRPT) algorithm, introduced by two of the authors of this paper [24]. It computes the hitting time for all $\binom{N}{2}$ pairs of points, the two rooted MSTs are grown in parallel simultaneously from each root, and it results in a pair of rooted MSTs that have the same number of edges at the hitting time. Building on the SDRPT concept we then define a modified algorithm, called the Dual Rooted Prim Tree (DRPT), that results in a pair of MSTs having different numbers of edges at the hitting time. Specifically, it selects a randomly chosen subset of pair of roots and grows MSTs sequentially and asymmetrically: at each stage of the Prim's algorithm, among the two new edges proposed for each MST only the rooted MST with the smallest edge is grown. Moreover, instead of using the "hitting time" as a dissimilarity measure, the length of the last constructed edge is used. This latter edge is a clique separator: its removal from the final graph disconnects the two rooted trees. The DRPT and the SDRPT have substantially different properties. In particular, the dissimilarity measure produced by the DRPT is a true metric regardless of the base dissimilarity measure used to define edge lengths for the Prim MST constructions.

Since the computation of the DRPT for all $\binom{N}{2}$ pair of vertices is necessary to construct a complete dissimilarity matrix, it may have a prohibitive computational cost. To address this point, we propose to create a consensus affinity matrix [33] based on the clusters produced by a subset of $M \ll \binom{N}{2}$ DRPT rooted at random pairs of points. As in the SDRPT of [24], or for consensus matrices in [53] spectral clustering to this matrix can then be used. Consensus clustering is a method for merging results from different algorithms, or from different clustering realizations associated with different initial conditions. This concept finds its origin in multi-classifier and multi-learner systems (see [23,33] for a brief history). The main idea is to empirically estimate performance by data partitioning, to create a set of clustering realizations that can be compared and combined [9,43,16]. The methods of cross-validation, bagging and boosting classifiers [10,18] are examples. We apply consensus clustering to dual rooted MSTs by applying it to the random selection of pairs of roots. As the proposed method accumulates evidence for clustering from each of the DRPTs, we refer to it as "Evidence Accumulating Clustering with Dual rooted Prim tree Cuts" (EAC-DC).

The EAC-DC approach has several features that we summarize here. First, the DRPT dissimilarity measure captures the dissimilarity of the MST neighborhoods of each pair of points. Second, since only a smaller random subset of pairs are used in the DRPT, it benefits from lower computational complexity than SDRPT with spectral clustering. We show by simulation and experiment that the EAC-DC outperforms state-of-the-art clustering methods on benchmark data sets. Third, as proven in the sequel, regardless of the base dissimilarity measure adopted to build the MST, the DRPT produces a dissimilarity measure which is a metric and this property can translate into improved performance relative to the SDRPT with spectral clustering. To illustrate this property, the DRPT is implemented on the symmetrized KL divergences between pairs of infrared star spectra to cluster stars in an astrophysical dataset.

The outline of the paper is as follows. Section 2 provides a brief introduction to minimal spanning trees and Prim's algorithm for general dissimilarity measures. In Section 2.2 dual rooted MST are discussed, the DRPT is proposed and its properties are discussed. Consensus clustering is applied to the DRPT dissimilarity measure in Section 3. Implementation and computational issues are also discussed in this section. Finally, after a brief review of clustering performance measures, an extensive comparative study is presented for both simulated and real datasets from the UCI repository of machine learning [2] and an astrophysical dataset for star classification.