Neurocomputing 151 (2015) 835–844

Contents lists available at ScienceDirect

Neurocomputing

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

A density-based similarity matrix construction for spectral clustering

Mario Beauchemin *

Natural Resources Canada,Canada Centre for Remote Sensing, 560 Rochester Street, Ottawa K1A 0E4, Canada

ARTICLE INFO

Article history:
Received 12 September 2013
Received in revised form
20 May 2014
Accepted 5 October 2014
Communicated by T. Heskes
Available online 17 October 2014

Keywords:
Affinity matrix
Non-parametric density estimation
Spectral clustering
K-means algorithm
Cluster ensembles

ABSTRACT

In the first part of this paper, we present a method to build affinity matrices for spectral clustering from a density estimator relying on K-means with subbagging procedure. The approach is anchored in the theoretical works of Wong (1980, 1982a, b) [13–15] on the asymptotic properties of K-means as a density estimation method. The subbagging procedure is introduced to improve the density estimate accuracy. The behavior of the proposed method is analyzed on diverse data sets and two new mechanisms are suggested to improve clustering results on non-convex data. In the second part of the paper, we establish a link between the presented method and the evidence accumulation clustering (EAC) approach by showing that a normalized version of the density-based similarity matrix is approximately equal to a normalized version of the co-association matrix. The co-association matrix provides the co-occurrence probability of data pairs assigned to a same cluster over multiple K-means clustering partitions. Experimental results on artificial and real data demonstrate the effectiveness of the method and provide empirical support for the established link.

1. Introduction

Over the last decade, spectral clustering has attracted a lot of attention for partitioning data. Part of its popularity comes from its good performance on data sets with non-convex clusters for which traditional algorithms such as K-means offers inferior performance. Spectral clustering has its root in graph partitioning problems and rely on the analysis of a similarity matrix[1]. In a recent paper, Zhang et al. [1] observed that (*i*) most studies on spectral clustering focus on the extraction of an optimal partition, given an affinity matrix and (*ii*) the problem of choosing an appropriate affinity matrix has received much less attention albeit its capital importance on the performance of spectral clustering is well-established [2–4]. The affinity matrix reflects the pairwise similarity relations among data points but its construction is a non-trivial task. Many ways of mapping a data set into an affinity matrix exist. The most common ones include ε-neighborhood graph, k-nearest neighbor graph and similarity functions [5]. The most widely used method is probably the pairwise similarity measure based on a Gaussian function $G(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\| \mathbf{x}_i - \mathbf{x}_j \|^2 / 2\sigma^2)$ where $\mathbf{x}$ denotes the data sample vector and σ is a free parameter. For the popular normalized-cut algorithm [6], it is suggested setting σ as a fraction, say 0.10, of the range of the pairwise distance encountered in the data set. However, Zelnik-Manor et al. [7] demonstrated that multiscale data sets cannot be properly partitioned with a unique value of σ and they proposed an algorithm to adapt a scaling measure according to a k-neighbor distance. Clearly, the k-neighbor distance relates to local density. Wang et al. [8] introduced clustering aggregation by probability accumulation where the co-association matrices built from K-means clustering are weighted by the average pairwise distance of each cluster. Recently, Zhang et al. [1] proposed a local density adaptive similarity measure based on neighborhood density information. The latter is based on the size of the shared neighborhood between two points.

The main focus of this paper is on density-based affinity matrix construction for spectral clustering. The idea of defining affinities from nonparametric density estimator was discussed in [9,10], where a link between graph-cut and kernel density estimation was established. It was shown that kernel density estimation could be used to define the similarity between two graph nodes. Subsequently, this probabilistic interpretation was thoroughly exploited in the mean shift spectral clustering algorithm with applications to image segmentation [11].

Kernel density estimation suffers, however, from the curse of dimensionality and presents mathematical challenges to establish their statistical characteristics in high dimension such as mean square error, consistency and rate of convergence [12]. Notably, multivariate density estimation from kernel estimation represents only one out of two historical prevailing approaches for density estimates, the other one being binned-type estimates in which the

* Tel.: +1 613 759 6449; fax: +1 613 759 6344.
E-mail address: mario.beauchemin@nrcan.gc.ca

[1] In this paper, we use the terms affinity, adjacency and similarity matrix interchangeably.

http://dx.doi.org/10.1016/j.neucom.2014.10.012
0925-2312/

histogram is the most well known representative. In this paper, we concentrate on binned-type estimates, which to the knowledge of the author have never received in-depth analysis for the construction of density-based affinity matrices. In the following, we present and test a locally adaptive affinity matrix construction based on K-means density estimation embedded within a subbagging procedure. Contrary to Parzen estimation, the proposed method is locally adaptive. The method is rooted in the works of Wong [13–15] on the asymptotic properties of K-means as a density estimation procedure and on the bootstrap-like proposal first explored in [16,17]. The algorithm is outlined in Algorithm 1. In brief, K-means is applied a large number of times P, each time with a number of clusters prescribed by the relationships derived in Wong [13–15] for appropriate density estimation (step 2). For each cluster within a given partition, the corresponding number of points and its associated support (volume) are computed (step 3). A density-based co-association matrix is then constructed on a pairwise basis where each pair within a cluster is assigned with the inverse support size of the cluster (step 6). Finally, the density-based affinity matrix is obtained from the average of the P density-based co-association matrices (step 8). This basic algorithm is further improved by the introduction of two new selection mechanisms to deal with violations of data distribution continuous assumption on density estimation and its impact on spectral clustering (shaded lines in Algorithm 1). The two proposed mechanisms resort on well-known statistical tests and prevent the presence of low quality partitions and bad clustering due to abrupt density changes.

Readers familiar with clustering ensemble methods have probably noticed the similarity in the steps described in Algorithm 1 with cluster ensemble construction. In the second part of the paper, we consolidate that similitude between both approaches. More precisely, we establish a direct connection between the density-based approach and the evidence accumulation clustering (EAC) paradigm introduced in Fred and Jain ([18], hereafter FJ). Such a connection helps explain the good performance (or failure) of EAC in relation to the characteristics of data sets. Furthermore, this connection permit avoiding the complexity level involved in the volume computation associated with the density-based affinity construction. To validate the connection, we first need to demonstrate theoretically and experimentally that density-based similarity construction based on K-means is a sound approach.

The paper is organized as follows. Section 2 provides the background material relevant to the proposed approach. Section 3 details the K-means-based density estimation method and Section 4 describes the construction of the density-based affinity matrix. The discontinuous distributions case is discussed in Section 5 and some algorithmic implementation details are given in Section 6. Section 7 presents the experimental results on both artificial and real data sets. The connection between the proposed method and the EAC is unveiled in Section 8. We conclude in Section 9.

2. Background works

2.1. K-means for density estimation

In a series of papers, Wong [13–15] demonstrate that, provided that the number of clusters k is of order $O([N/\log N]^{1/3})$, K-means can be used to construct a uniformly consistent histogram estimate of an unknown density, $f(\mathbf{x})$. Using the asymptotic properties of K-means, Wong [13] showed that the sizes of K-means cluster intervals for continuous univariate distributions are proportional to $f(\mathbf{x})^{-1/3}$ at the midpoints of the intervals and therefore are adaptive to the underlying density, $f(\mathbf{x})$. This implies that the cluster size will be large where data are sparse and small where data are dense. This behavior is a highly desirable property for a density estimator [17,19]. For multivariate distributions, Wong [14] proposes to estimate the density according to $f(\bar{\mathbf{x}}_i) = c\ N_i^{1+d/2}\ WSS_i^{-d/2}$, where $\bar{\mathbf{x}}_i$ and WSS_i are the sample mean and within-cluster sum of squares of the ith cluster containing N_i objects, c is a proportionality constant, and d is the data dimensionality ($d>1$). The estimate, $f(\bar{\mathbf{x}}_i)$, is based on the fact that the volume of the i^{th} cluster is approximated by $[N_i^{-1}WSS_i]^{d/2}$. It is expected that such volume estimation be valid for low dimensions. The case for high dimensions will be discussed in Section 6.2. Wong suggested two empirical expressions to determine the number k of cluster: $k_{W1}=4N^{0.3}$ [15] and $k_{W2}=7(N/\log N)^{1/3}$ [14]. These two expressions are of key practical value for the proposed method. In general, almost-sure L_1 consistency of density estimate derived from data-driven partitions has been demonstrated in [20].

2.2. Bootstrap aggregation (bagging) for density estimate

Generally, the K-means algorithm is initialized from seed points selected randomly among the data set and they represents typically only a few percent of the data set. It is well known that K-means results are sensitive to initialization. Such an algorithm is therefore unstable in the sense that perturbations, i.e. different initializations, result in different outputs. Browne [21], in a way similar to [16,17,22,23,12] has utilized subbagging to stabilize density estimate from random tessellation. Subbagging is similar to bagging except that only a subset of the data set is utilized [24]. The idea explored by the aforementioned authors consists in using a random subsample of the data set to construct Delauney tessellation (or its dual, the Voronoi tessellation). Like in histogram-based density estimate, the number of data set points in a tile divided by its support size provides a local density estimate. By additional re-sampling (bootstrap) an average density over tiles is obtained (aggregation). Champaneri [12] proved that binning based on Delaunay tessellation is consistent, conditionally maximum likelihood and has asymptotic distribution. For random tessellation, however, and contrary to Wong's works, there is no *a priori* prescription to find the appropriate subset size for density estimation from random tessellation. The latter observation was decisive to us in considering K-means tessellation instead of a random one. Although Browne [21] proposed an *a posteriori* method to determine the appropriate size using Akaike information criterion, this requires computing density estimate for a large number of sub-sample sizes. Notice that we do not discard random tessellation as a valid alternative.

3. Improved density estimate: K-means with subbagging

Considering the previous observations, one expects that density estimation from K-means should benefit from procedures such as bootstrap aggregating. In particular, subbagging is well adapted for Wong's approach as the method intrinsically uses random subsampling. The proposed K-means algorithm for density estimate with subbagging works as follows:

1. Choose $k\sim O([N/\log N]^{1/3})$, predefine a large number of partitions P;
2. Apply K-means P times with k random cluster centers selected from the full sample;
3. For each generated partition t, $t=1$ to P, assign to every $\mathbf{x}_i$, $i=1$ to N, its corresponding cluster point density estimate δ_i^t. The latter is given by $N_i^t/(NS_i^t)$ where N_i^t is the number of points in the cluster to which $\mathbf{x}_i$ belong to and S_i^t is the cluster support (volume);
4. The density estimate $\hat{\delta}_i$ at $\mathbf{x}_i$ is the average density evaluated overall partitions: $\hat{\delta}_i = P^{-1}\sum_{t=1}^{P}\delta_i^t$.